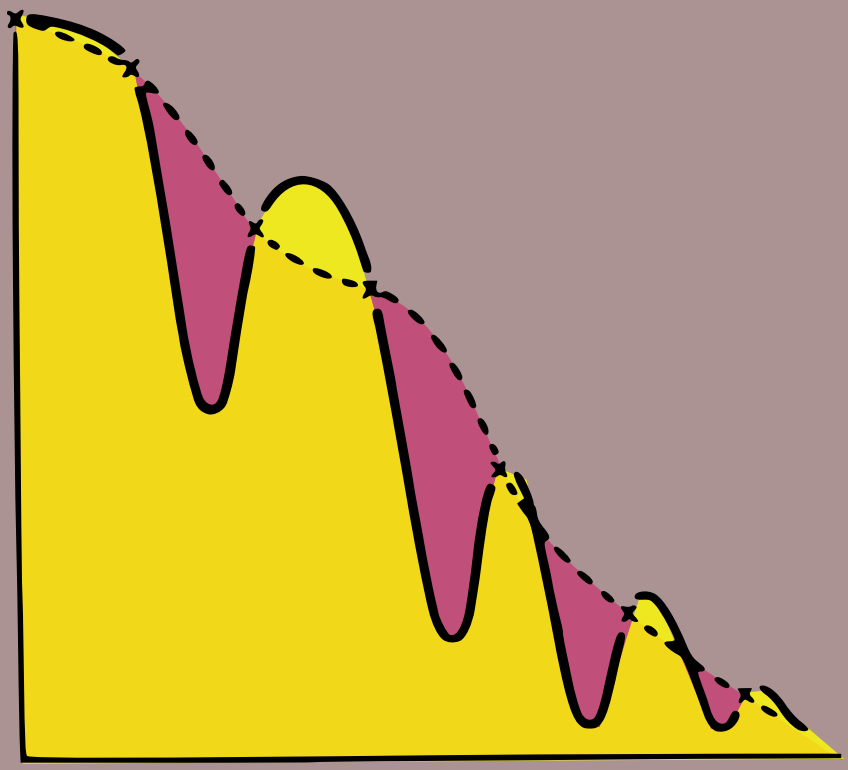
N. Bakhvalov

MÉTHODES NUMÉRIQUES

Analyse, algèbre, équations différentielles ordinaires



Éditions Mir Moscou

Н. С. БАХВАЛОВ

ЧИСЛЕННЫЕ МЕТОДЫ

АНАЛИЗ, АЛГЕБРА, ОБЫКНОВЕННЫЕ ДИФФЕРЕНЦИАЛЬНЫЕ УРАВНЕНИЯ

ИЗДАТЕЛЬСТВО «НАУКА»
МОСКВА

N. BAKHVALOV

MÉTHODES NUMÉRIQUES

Analyse, algèbre, équations différentielles ordinaires

Deuxième édition

ÉDITIONS MIR • MOSCOU

Traduit du russe
par IRINA PÉTROVA

На французском языке

© Издательство «Наука» 1973

© Traduction francaise Editions Mir 1976

TABLE DES MATIÈRES

Avant-propos 9
Introduction 11

Première partie
MÉTHODES NUMÉRIQUES DE L'ANALYSE MATHÉMATIQUE

Chapitre premier. ERREURS 17
§ 1. Sources et classification des erreurs 17
§ 2. Enregistrement des nombres sur machines à calculer électroniques 20
§ 3. Erreur absolue et erreur relative. Ecritures des données 21
§ 4. Erreur de calcul 23
§ 5. Erreur sur une fonction 24

Chapitre II. INTERPOLATION ET PROBLÈMES CONNEXES 30
§ 1. Approximation des fonctions. Position du problème 31
§ 2. Polynôme d'interpolation de Lagrange 34
§ 3. Evaluation du reste du polynôme d'interpolation de Lagrange 36
§ 4. Différences divisées et leurs propriétés 37
§ 5. Formule d'interpolation de Newton par différences divisées . . 39
§ 6. Différences divisées et interpolation basée sur des points multiples 42
§ 7. Equations aux différences finies 46
§ 8. Polynômes de Tchebycheff 55
§ 9. Minimisation de l'évaluation du reste d'une formule d'interpolation 59
§ 10. Différences finies 61
§ 11. Formules d'interpolation de Newton pour des points équidistants 65
§ 12. Formules d'interpolation de Bessel et d'Everett. Formation des tableaux 66
§ 13. Erreur d'arrondi dans l'interpolation 75
§ 14. Applications de l'interpolation. Interpolation inverse 77
§ 15. Systèmes orthogonaux et leurs propriétés 78
§ 16. Polynômes orthogonaux 84
§ 17. Dérivation numérique 88
§ 18. Erreur de calcul dans les formules de dérivation numérique . . . 91

Chapitre III. INTÉGRATION NUMÉRIQUE 94
§ 1. Formules de quadrature de Newton-Cotes 94
§ 2. Evaluation de l'erreur dans une formule de quadrature sur une classe de fonctions 102
§ 3. Formules de quadrature de Gauss 106
§ 4. Evaluation pratique de l'erreur dans des formules de quadrature élémentaires 117

§ 5. Intégration de fonctions fortement oscillantes 122
§ 6. Amélioration de la précision d'intégration par partage équidistant du segment 125
§ 7. Position des problèmes d'optimisation 131
§ 8. Formules de quadrature optimales sur des classes de fonctions à une dérivée 135
§ 9. Optimisation de la répartition des points de base des formules de quadrature 142
§ 10. Exemples d'optimisation de la répartition des points d'intégration 148
§ 11. Partie principale de l'erreur 154
§ 12. Formules d'Euler et de Gregory 158
§ 13. Règle d'évaluation pratique de l'erreur de Runge 161
§ 14. Formules de Romberg 168
§ 15. Expériences et leur discussion 171
§ 16. Calcul d'intégrales dans des cas non réguliers 179
§ 17. Principes d'établissement des programmes usuels avec recherche automatique du pas 185
§ 18. Programmes usuels d'intégration numérique 192

Chapitre IV. APPROXIMATION DES FONCTIONS ET PROBLÈMES CONNEXES 201

§ 1. Meilleures approximations dans un espace vectoriel normé . . . 201
§ 2. Meilleure approximation dans un espace de Hilbert et sa construction 203
§ 3. Transformation de Fourier discrète 208
§ 4. Transformation de Fourier rapide 212
§ 5. Meilleure approximation uniforme 215
§ 6. Exemples de meilleure approximation uniforme 219
§ 7. Méthode itérative de construction du polynôme de meilleure approximation uniforme 225
§ 8. Ecriture du polynôme 232
§ 9. Procédés de calcul des fonctions élémentaires 239
§ 10. Vitesse d'approximation des fonctions de différentes classes . . 243
§ 11. Interpolation et approximation par des fonctions-spline 246
§ 12. Entropie et ε-entropie 252

Chapitre V. PROBLÈMES À PLUSIEURS DIMENSIONS 259

§ 1. Méthode des coefficients indéterminés 260
§ 2. Méthode des moindres carrés 261
§ 3. Méthode de régularisation 263
§ 4. Exemple de régularisation 264
§ 5. Réduction des problèmes multidimensionnels aux problèmes unidimensionnels 270
§ 6. Evaluation de l'erreur dans l'intégration numérique sur un réseau régulier 278
§ 7. Minoration de l'erreur d'intégration numérique 281
§ 8. Optimisation de l'estimation de l'erreur sur des classes plus vastes de procédés d'intégration 284
§ 9. Méthode de Monte-Carlo 289
§ 10. Légitimité des méthodes indéterministes de résolution des problèmes 293
§ 11. Accélération de la convergence de la méthode de Monte-Carlo . . 296
§ 12. Formules de quadrature de précision élevée aux points aléatoires 299
§ 13. Choix de la méthode 304

Deuxième partie
PROBLÈMES D'ALGÈBRE ET D'OPTIMISATION

Chapitre VI. MÉTHODES NUMÉRIQUES DE L'ALGÈBRE 311

§ 1. Méthodes d'élimination successive des inconnues 312
§ 2. Méthode d'orthogonalisation 320
§ 3. Méthode itérative simple 322
§ 4. Processus itératif réel 327
§ 5. Spectre d'une famille de matrices 330
§ 6. Procédé δ^2 d'évaluation pratique de l'erreur et d'accélération de la convergence 336
§ 7. Optimisation de la vitesse de convergence des processus itératifs 339
§ 8. Méthode de Seidel 349
§ 9. Méthode de la plus grande pente 355
§ 10. Méthode du gradient conjugué 358
§ 11. Méthode de Monte-Carlo pour des systèmes d'équations linéaires 364
§ 12. Méthodes par itérations avec utilisation des opérateurs spectralement équivalents 371
§ 13. Erreur sur la solution approchée d'un système d'équations et conditionnement des matrices. Régularisation 374
§ 14. Problème de valeurs propres 380
§ 15. Résolution du problème global de valeurs propres d'une matrice symétrique par la méthode des rotations 385

Chapitre VII. RÉSOLUTION DES SYSTÈMES D'ÉQUATIONS NON LINÉAIRES ET DES PROBLÈMES D'OPTIMISATION 390

§ 1. Méthode itérative simple et problèmes connexes 391
§ 2. Méthode de Newton pour les systèmes d'équations non linéaires 396
§ 3. Autres méthodes de résolution d'une équation 401
§ 4. Méthodes de descente 405
§ 5. Autres procédés de réduction de la dimension des problèmes . . 410
§ 6. Méthodes de stationnarisation 414
§ 7. Que faut-il optimiser? 420
§ 8. Comment optimiser? 425

Troisième partie
MÉTHODES DE RÉSOLUTION NUMÉRIQUE DES ÉQUATIONS DIFFÉRENTIELLES ORDINAIRES

Chapitre VIII. MÉTHODES DE RÉSOLUTION DU PROBLÈME DE CAUCHY 431

§ 1. Développement de la solution en série de Taylor 432
§ 2. Méthodes de Runge-Kutta 434
§ 3. Méthodes avec contrôle de l'erreur par pas 442
§ 4. Evaluation d'erreur dans les méthodes à pas séparés 443
§ 5. Méthodes des différences finies 449
§ 6. Méthode des coefficients indéterminés 453
§ 7. Etude des propriétés des méthodes des différences finies sur des problèmes modèles 458
§ 8. Évaluation d'erreur dans les méthodes des différences finies . . 485

§ 9. Partie principale de l'erreur 469
§ 10. Etude des propriétés des méthodes des différences finies sur des modèles plus précis 474
§ 11. Intégration de systèmes d'équations 483
§ 12. Quelques questions générales 492
§ 13. Formules d'intégration numérique des équations du deuxième ordre 499
§ 14. Evaluation de l'erreur dans la résolution numérique du problème de Cauchy pour une équation du deuxième ordre 503
§ 15. Méthodes bilatérales 508

Chapitre IX. MÉTHODES DE RÉSOLUTION DES PROBLÈMES AUX LIMITES POUR LES ÉQUATIONS DIFFÉRENTIELLES ORDINAIRES 515
§ 1. Méthodes simples de résolution du problème aux limites pour une équation du deuxième ordre 515
§ 2. Fonction de Green du problème aux limites discrétisé 522
§ 3. Résolution d'un simple problème aux limites discrétisé 527
§ 4. Fermetures des algorithmes de calcul 536
§ 5. Position des problèmes aux limites pour des systèmes linéaires du premier ordre 543
§ 6. Algorithmes de résolution des problèmes aux limites pour des systèmes d'équations du premier ordre 549
§ 7. Méthodes du balayage orthogonal différentiel 555
§ 8. Problèmes aux limites non linéaires 560
§ 9. Approximations du type spécial 569
§ 10. Recherche des valeurs propres par les méthodes des différences finies 577
§ 11. Optimisation de la répartition des points d'intégration 582
§ 12. Influence de l'erreur de calcul en fonction de l'écriture de l'équation aux différences finies 587
§ 13. Evaluation de l'erreur de calcul pour le problème aux limites résolu par la méthode du balayage 593

Bibliographie 597
Index 603

AVANT-PROPOS

Le présent livre est issu du cours que l'auteur donne depuis dix ans à la Faculté physico-mathématique et à la Faculté d'analyse numérique et de cybernétique de l'Université de Moscou ainsi que des travaux du Centre de calcul de ladite Université.

Comme tout manuel de méthodes numériques, cet ouvrage expose les notions fondamentales de la théorie concernant, dans notre cas, l'approximation de fonctions, l'intégration, les problèmes d'algèbre et d'optimisation, la résolution d'équations différentielles.

Nous assistons aujourd'hui à un développement impétueux du matériel de calcul (la valeur du parc augmente de plusieurs ordres par décennie). Les machines et les méthodes numériques sont constamment sollicitées afin de faire face à des problèmes toujours plus variés: D'où une reconversion permanente de nos idées sur l'ensemble des questions liées à l'application des ordinateurs et les tâches que nous imposons aux procédés de calcul. Pour toutes ces raisons donc il est impossible de proposer un manuel panacée, et quand on choisit un mode opératoire susceptible de nous donner la solution de tel problème compliqué, tout manuel n'est souvent qu'une sorte de guide général dont la seule vocation est d'inspirer le chercheur.

La création et la mise en pratique des méthodes numériques relèvent entre autres de l'étude purement théorique des techniques de résolution des problèmes types, de l'analyse du travail de l'algorithme dans le cas des problèmes modèles et de l'expérience numérique. On ne saurait surestimer le rôle du choix de la direction des recherches de la construction de modèles mathématiques des phénomènes étudiés, de l'établissement des contacts avec les représentants d'autres disciplines.

Les facteurs énumérés se particularisent suivant les cas, si bien qu'il est difficile de donner une recette générale. Le lecteur est donc prié de voir dans les raisonnements de l'auteur non pas un guide pour l'action mais un point de vue parmi tant d'autres.

L'efficacité des ordinateurs et des méthodes numériques se mesure au nombre de problèmes ardus résolus ainsi qu'au nombre d'utilisateurs gagnés en augmentant le « confort » de leur accès à la machine. Ce dernier but exige qu'on rédige des programmes standard de résolution de problèmes mathématiques types. Aussi le présent livre

réserve-t-il certaine place aux questions de la théorie des méthodes numériques qui surgissent lors de l'élaboration de tels programmes.

Cet ouvrage doit beaucoup à Monsieur le Professeur S. Bakhvalov, père de l'auteur, à Messieurs A. Kolmogorov, S. Sobolev, A. Tikhonov, de l'Académie des Sciences de l'U.R.S.S., à Monsieur N. Tchétaev, membre correspondant de l'Académie, à Messieurs les Professeurs I. Bérézine, S. Godounov, à Messieurs N. Jidkov, V. Karmanov, A. Filippov, chargés de cours, qui ont notablement influencé les idées de l'auteur sur les questions abordées.

L'auteur adresse ses remerciements à Monsieur le Professeur E. Volkov, qui, chargé de recenser le présent livre, lui a apporté de très précieux conseils.

La gratitude de l'auteur va à Messieurs les Professeurs V. Voïévodine, N. Kouznétsov, à Messieurs F. Vassiliev, E. Diakonov, I. Jiléikine, V. Morozov, chargés de cours, chez qui il a trouvé d'utiles suggestions qui ont eu pour effet d'améliorer le manuscrit.

L'auteur est fort reconnaissant à Madame R. Razouméiko et à Messieurs B. Bérézine, G. Kobelkov, A. Koukarkine, K. Ossipenko, qui l'ont grandement aidé à rédiger le manuscrit.

Nikolaï Bakhvalov

INTRODUCTION

Avant d'exposer la théorie des méthodes numériques, essayons de la situer par rapport aux autres branches scientifiques et de donner un bref aperçu des problèmes que posent ses applications.

Les mathématiques devaient primitivement faire face à des problèmes pratiques tels que l'arpentage, la navigation, etc. D'où leur caractère numérique : leur but était l'obtention d'une solution sous forme de nombres.

De tout temps les mathématiciens s'intéressaient à la résolution numérique des problèmes. Les plus grands savants d'antan étudiaient les phénomènes de la nature en les décrivant mathématiquement (en construisant un modèle mathématique), puis en analysant cette description. L'étude de modèles plus compliqués donna naissance à des méthodes numériques spéciales dues notamment à Newton, Euler, Lobatchevski, Gauss, Tchebycheff, Hermite...

Nous assistons aujourd'hui à une extension brusque du champ d'application des mathématiques, extension qui s'explique surtout par l'avènement et le développement du matériel de calcul. Avec la création des calculateurs à programme enregistré, la vitesse des opérations arithmétiques passa en moins de trente ans de 0,1 opération par seconde (calcul à la main) à 3 000 000 opérations (ordinateurs modernes de série), soit un accroissement d'environ $5\,000^2$ fois.

Les machines à calculer nous ouvrent des horizons nouveaux, ce qui autorise certains à évoquer la révolution industrielle provoquée par l'invention de la machine à vapeur. Rappelons cependant qu'à la suite de ce bouleversement technique et du développement ultérieur, deux siècles durant, des sciences et de la technique, la vitesse de déplacement de l'homme est passée de 6 kilomètres à l'heure (marche à pied) à 30 000 km/h avec les voyages interplanétaires, soit un accroissement de 5 000 fois.

L'opinion répandue sur l'omnipotence des ordinateurs donne l'impression que les mathématiciens n'éprouvent aucune difficulté, ou peu s'en faut, à résoudre numériquement un problème et n'ont que faire de nouvelles méthodes de calcul. Or, rien n'est plus faux.

Une des lois présidant au progrès des sciences est que la société pour se développer leur assigne souvent des tâches au-dessus de leurs moyens. La généralisation des mathématiques appliquées

a conditionné la mathématisation de sciences telles que la chimie, l'économie, la biologie, la géologie, la géographie, la psychologie, la médecine et certaines techniques. Il s'agit en l'occurrence de construire les modèles mathématiques de processus et phénomènes et de mettre au point les méthodes d'étude adéquates.

D'autre part, la simulation mathématique et l'étude des modèles afin d'expliquer le présent et le futur des phénomènes sont dans la tradition des disciplines telles que la physique et la mécanique. Ainsi, la découverte de Neptune offre un exemple suggestif d'étude des modèles mathématiques alliée au traitement des résultats des observations bien avant l'apparition des machines à calculer.

Néanmoins, on ne progressait que lentement. En effet, on n'arrivait pas toujours à résoudre les problèmes qui se posaient et on se contentait des modèles les plus simples. L'utilisation des machines à calculer électroniques et la généralisation de l'enseignement des mathématiques accrurent les possibilités de construction et d'étude des modèles. De plus en plus souvent, les résultats des calculs permettent de déceler et de prédire des phénomènes sans précédent, si bien qu'on peut parler d'expérience mathématique. Dans certains domaines de la recherche le calcul numérique occupe une place tellement importante que si on constate un écart entre les résultats des calculs et ceux des expériences, ce sont ces derniers qu'on teste afin de détecter une erreur.

On s'imagine mal comment la recherche atomique, l'investigation du cosmos et l'économie auraient pu progresser sans l'apport des machines à calculer et des méthodes numériques.

Le besoin de résoudre numériquement de nombreux problèmes a donné naissance à de nombreux procédés. De plus, voilà trente ans qu'on interprète intensément sur le plan théorique les techniques anciennes et nouvelles, qu'on essaye de les systématiser.

En calcul scientifique les efforts sont centrés sur l'élaboration des méthodes de résolution des problèmes mathématiques types. On entend par là des problèmes d'analyse (approximation, dérivation, intégration), d'algèbre, la résolution d'équations différentielles et intégrales, les problèmes d'optimisation, les problèmes inverses (ces deux dernières directions de la théorie des méthodes numériques ne datent que d'une quinzaine d'années). Ces recherches qui sont d'un grand secours lorsqu'on affronte des problèmes concrets, contribuent largement à élargir le champ d'application des ordinateurs et des mathématiques en général.

Considérons un exemple suggestif. La résolution d'équations aux dérivées partielles se ramène à celle d'un système d'équations linéaires avec une matrice dont chaque ligne contient 5 à 10 éléments non nuls. Tant que les machines mathématiques numériques n'existaient pas, le nombre des inconnues ne devait pas dépasser 10^2. Or, on résout aujourd'hui des systèmes dont les inconnues sont de 10^5 à 10^6. Si l'on

se servait des méthodes vieilles d'une trentaine d'années, on se bornerait à des équations telles que le nombre d'inconnues serait compris entre 10^3 et 10^4 (le temps machine restant le même).

L'irruption brusque des mathématiques dans de nombreuses branches de la connaissance tient à ce que les mêmes modèles et méthodes mathématiques s'appliquent à nombre de phénomènes d'une structure formelle analogue. Le modèle mathématique d'un processus quelconque se forme souvent au cours de l'étude d'autres phénomènes ou des échafaudages mathématiques abstraits bien avant qu'on aborde le phénomène en question. En particulier, la théorie des méthodes numériques, tout comme les mathématiques pures, gagne beaucoup à élaborer des constructions générales. Le problème est néanmoins approché différemment. Pour le mathématicien « pur », résoudre un problème c'est démontrer l'existence d'une solution et proposer un processus convergent vers celle-ci. Les résultats sont par eux-mêmes utiles au mathématicien « appliqué », qui veut de plus que l'approximation ne coûte trop en temps machine et en occupation mémoire. La convergence du processus l'intéresse certes, mais il désire également en connaître l'ordre. La résolution numérique des problèmes suscite également des questions liées à la stabilité du résultat par rapport aux perturbations des données initiales et aux arrondis.

Des disciplines mathématiques autres que la théorie des méthodes numériques, qui sont tributaires elles aussi des machines à calculer électroniques, connaissent un essor prodigieux. L'application du calcul numérique et des machines dans les sciences de la nature s'est répercutée sur les branches mathématiques traditionnelles. Ainsi, l'étude des systèmes hyperboliques quasi linéaires doit beaucoup au calcul mathématique.

Science de la nature, les mathématiques évoluent comme telles, et leur développement fut déterminé, de longues années durant, par les besoins de la mécanique et de la physique. En mathématisant telles disciplines on doit s'attendre à une action en retour de leur part qui ne manquera pas de changer radicalement la physionomie des mathématiques.

Et voici une autre raison de l'influence de l'aspect application des mathématiques sur leur aspect théorie. Les effets de la recherche appliquée sont immédiats, ce qui accroît le prestige de la science, la rend moins occulte aux yeux de la société et incite à augmenter les investissements pour favoriser son développement.

Le mathématicien « appliqué » se heurte à toutes sortes de difficultés de nature pas nécessairement mathématique.

Sans prétendre substituer des préceptes à l'expérience personnelle, nous voudrions toutefois attirer votre attention, cher Lecteur, sur certaines questions de caractère général. La systématisation ci-dessous est fortuite et conventionnelle, mais elle n'est ni plus bonne ni

plus mauvaise que de nombreuses autres qui ont elles aussi droit de cité.

1. Le choix de la direction des recherches a une importance capitale. A vrai dire, la liberté d'action du chercheur est plutôt réduite en l'occurrence car il travaille en général sur un canevas imposé du dehors.

Ceci étant, n'oublions pas que les problèmes peuvent être 1) simples, 2) difficiles, 3) très difficiles. Laissez les premiers : ils seront résolus progressivement sans votre intervention. Nous doutons fort que le groupe 3) soit à votre taille. Occupez-vous donc des problèmes dits « difficiles ».

2. Il faut savoir formuler dans le langage mathématique des problèmes concrets de physique, mécanique, économie, des problèmes de l'ingénieur, c'est-à-dire mettre en équation le phénomène étudié.

Le mathématicien néophyte se plaint souvent des représentants des autres sciences en contact avec lui, qui ne savent « même » pas formuler leurs problèmes. Formuler correctement un problème est tout aussi difficile que le résoudre et n'allez pas croire que quelqu'un s'en chargera à votre place. En posant un problème, centrez votre attention sur le but des recherches ; le modèle mathématique adopté n'est pas une chose figée, liée une fois pour toutes avec le phénomène en question, il est fonction de l'objectif que vous poursuivez. Avant d'écrire des équations différentielles, de choisir la méthode de résolution et de s'approcher de l'ordinateur, réfléchissez si le calcul vaudra la peine d'être fait. En même temps, il faut vous faire à l'idée que la plupart des résultats seront rejetés sitôt obtenus. En effet, en calcul scientifique, il est difficile de prédire la nature et la forme du résultat éventuel, d'indiquer la voie dans laquelle on cherchera la solution numérique. Le but de l'étude et la description du problème sont d'habitude précisés dans le cadre de contacts entre les représentants de telles sciences et organisations (clients) et les mathématiciens (chercheurs ou exécutants).

3. Un travail fructueux dans les sciences appliquées suppose une bonne formation mathématique qui seule est capable d'adapter le chercheur à une variation incessante de problèmes à résoudre. L'un des arguments en faveur de la connaissance des branches « inutiles » de prime abord au praticien est qu'elles permettent un maniement plus sûr et plus aisé des branches « utiles ».

4. Une connaissance parfaite des mathématiques, des méthodes numériques et une pratique suffisante des machines à calculer électroniques ne garantissent nullement la résolution immédiate d'un problème de mathématiques appliquées. Souvent une *mise au point* des procédés s'impose afin de les adapter à tel ou tel problème concret. Un cas typique est celui où l'on se sert de méthodes non justifiables théoriquement et où les estimations théoriques de l'erreur de

méthode sont inacceptables pour un praticien. En choisissant la méthode ou en analysant les résultats, on doit se guider sur l'expérience, l'intuition et la comparaison des tests ; ce faisant on doit garantir la certitude des résultats. Aussi n'aboutira-t-on qu'en s'arrachant au formalisme et en raisonnant par analogie, politique qui permet de garantir l'exactitude du résultat dans les cas où, du point de vue logique et mathématique, on ne peut être sûr de rien.

Considérons l'autre aspect de la question. En résolvant numériquement des problèmes ardus, extérieurs à sa spécialité, le mathématicien procède souvent comme un naturaliste en se fondant sur son expérience et sur des raisonnements « vraisemblables ». Il serait bon que ce travail empirique s'appuie sur l'élaboration théorique des méthodes, sur une vérification scrupuleuse de leur qualité par des problèmes de contrôle de solution connue, enfin sur une confrontation objective avec les expériences. Si l'on progresse longtemps dans une direction sans posséder de pareilles arrière-gardes, on risque de perdre le sens de la perspective, de commencer à douter des résultats qu'on est en train d'obtenir. On dit qu'un bon théoricien interprète comme bon lui semble n'importe quel résultat des calculs et des expériences.

5. Une fois les calculs terminés, il faut utiliser les résultats dans la pratique. A vrai dire, on s'y prépare dès l'instant où l'on pose le problème si bien qu'il y a imbrication des diverses phases de la résolution et de la mise en pratique ; quand ils formulent le problème et le résolvent, le client et l'exécutant précisent ensemble la position du problème et partant son champ d'application. Les mathématiques associées aux ordinateurs s'implantent dans les domaines les plus divers et il arrive souvent que les clients n'ont aucune expérience du calcul automatique. Avant toute chose il importe de surmonter leur appréhension des mathématiques qui s'imposent impérieusement ; le client ne mettra à profit les résultats des calculs que s'il les interprète de son point de vue et s'il se convainc qu'ils peuvent et doivent être utilisés. Sous réserve de bons contacts entre le client et le mathématicien, le premier comprend que les mathématiques et les ordinateurs satisfont une bonne partie de ses besoins mais pas tous, tandis que le second se rend compte qu'il a fait quelque chose mais pas tout ce qu'il fallait pour résoudre réellement le problème posé.

Qu'on se rappelle l'histoire de la découverte de Neptune. Un an avant Le Verrier, Adams découvre cette planète par le calcul, mais, faute de contacts avec les astronomes, son résultat n'est pas vérifié expérimentalement.

6. Il est essentiel, pour le mathématicien « appliqué », de respecter les délais fixés.

Le client est souvent enjoint de prendre la décision à une date précise. Si les recherches traînent, la décision est prise quand même,

mais à partir de considérations plus grossières, empiriques, voire tout simplement à la légère. Et le mathématicien risque de perdre à tout jamais la confiance du client.

Aussi vaut-il mieux trouver une solution satisfaisante dans les délais prévus plutôt qu'une solution complète au moment où on n'en a plus besoin.

Bibliographie de l'introduction

I [15], II [40 41 43]

PREMIÈRE PARTIE

MÉTHODES NUMÉRIQUES DE L'ANALYSE MATHÉMATIQUE

CHAPITRE PREMIER

ERREURS

§ 1. Sources et classification des erreurs

On distingue trois sortes d'erreurs commises en résolvant un problème :

1. Des erreurs de description mathématique, y compris des erreurs de données.

2. Une précision insuffisante de la méthode utilisée : la solution exacte du problème posé nécessite des opérations arithmétiques en nombre très grand, voire illimité, ce qui implique des approximations.

3. Des erreurs d'arrondi sur les données d'entrée, les opérandes et les données de sortie.

Nous allons donner à ces erreurs respectivement les noms de

1) *erreurs inhérentes*,

2) *erreurs de méthode*,

3) *erreurs de calcul*.

Illustrons ces définitions. Soit un pendule (fig. 1.1.1) *) qui se met en mouvement à l'instant $t = t_0$. Quelle sera sa position à l'instant t_1?

L'équation différentielle décrivant l'oscillation du pendule s'écrit

$$l \frac{d^2\varphi}{dt^2} + g \sin \varphi + \mu \frac{d\varphi}{dt} = 0, \qquad (1)$$

l étant la longueur du pendule, g l'accélération de la pesanteur, μ le coefficient de frottement.

Cette loi étant admise, la solution du problème acquiert une erreur inhérente ne serait-ce que parce que, en réalité, la relation

*) Les formules et les figures sont référencées dans le texte par trois nombres désignant respectivement le chapitre, le paragraphe et le numéro de la figure ou de la formule; par deux nombres qui sont le paragraphe et le numéro (dans les limites d'un même chapitre); par un seul nombre s'il s'agit d'un même paragraphe.

entre le frottement et la vitesse n'est pas tout à fait linéaire; une autre source d'erreur inhérente est la définition de l, g, μ, t_0, $\varphi(t_0)$, $\varphi'(t_0)$.

Cette erreur inévitable est hors calcul, et seules une description plus rigoureuse du phénomène physique et une définition plus précise des paramètres permettent de la réduire. L'équation différentielle (1) n'est pas résoluble sous une forme explicite. Sa résolution exige le recours à une méthode numérique qui donne lieu à une erreur supplémentaire.

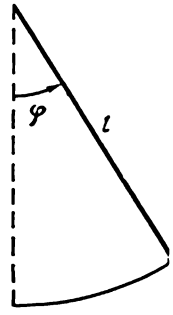


Fig. 1.1.1

L'erreur de calcul peut résulter par exemple de la longueur finie des nombres avec lesquels on opère.

Introduisons les définitions formelles.

Soient I la valeur exacte du paramètre cherché (dans notre cas, l'angle φ dont on dévie le pendule de la verticale à l'instant t_1), $\tilde{I}$ la valeur du paramètre correspondant à la description mathématique adoptée (i.e. la valeur $\varphi(t_1)$ de la solution de l'équation (1)), $\tilde{I}_h$ la solution obtenue par la méthode numérique sans arrondi, $\tilde{I}_h^*$ une solution approchée obtenue par des calculs réels. Alors

$$\begin{aligned} \rho_1 &= \tilde{I} - I \quad \text{est une erreur inhérente,} \\ \rho_2 &= \tilde{I}_h - \tilde{I} \quad \text{une erreur de méthode,} \\ \rho_3 &= \tilde{I}_h^* - \tilde{I}_h \quad \text{une erreur de calcul.} \end{aligned} \tag{2}$$

L'erreur totale $\rho_0 = \tilde{I}_h^* - I$, i.e. la différence entre la solution calculée et la solution exacte, vérifie l'égalité évidente

$$\rho_0 = \rho_1 + \rho_2 + \rho_3. \tag{3}$$

Dans de nombreux cas, on entend par « erreur » non pas les différences ci-dessus mais des mesures de proximité des valeurs approchées. Ainsi, on pose, pour les scalaires:

$$\rho_0 = |\tilde{I}_h^* - I|, \ \rho_1 = |\tilde{I} - I|, \ \rho_2 = |\tilde{I}_h - \tilde{I}|, \ \rho_3 = |\tilde{I}_h^* - \tilde{I}_h|;$$

avec ces notations, on a au lieu de (3)

$$\rho_0 \leqslant \rho_1 + \rho_2 + \rho_3. \tag{4}$$

Dans d'autres cas, la solution I et les approximations $\tilde{I}$, $\tilde{I}_h$, $\tilde{I}_h^*$ sont éléments d'espaces fonctionnels souvent distincts. $\tilde{I}$ peut appartenir à l'espace F des fonctions continues sur $[0, 1]$, et $\tilde{I}_h$ à l'espace F_h des fonctions discrètes f_h définies aux points $x_n = nh$, $n = 0, \ldots, h^{-1}$, avec h^{-1} entier. On mesure alors l'erreur par un degré de proximité $\rho(z_1, z_2)$, z_1 et z_2 appartenant à un même sous-espace ou à des sous-espaces différents. $\rho(z_1, z_2)$ est astreint d'être la mesure naturelle de l'erreur et de vérifier l'inégalité triangulaire

$$\rho(z_1, z_3) \leqslant \rho(z_1, z_2) + \rho(z_2, z_3) \tag{5}$$

quels que soient z_1, z_2, $z_3 \in F$, F_h. Ceci étant, $\rho(z_1, z_2) = 0$ n'implique pas forcément $z_1 = z_2$ si bien que la fonction $\rho(z_1, z_2)$ n'est pas nécessairement la distance dans un certain espace métrique.

Dans l'exemple considéré on peut poser

$$\rho(f_1, f_2) = \max_{0 \leq n \leq h^{-1}} |f_1(nh) - f_2(nh)|$$

quels que soient les espaces contenant f_1 et f_2.

Les mesures d'erreurs sont définies respectivement par les égalités

$$\rho_0 = \rho(\widetilde{I}_h^*, I),\ \rho_1 = \rho(\widetilde{I}, I),\ \rho_2 = \rho(\widetilde{I}_h, \widetilde{I}),\ \rho_3 = \rho(\widetilde{I}_h^*, \widetilde{I}_h);$$

compte tenu de (5), elles satisfont à (4).

Une question peut se poser: pourquoi étudier l'erreur inhérente puisque à priori elle est inévitable? Cette question paraît naturelle venant du mathématicien à qui l'on propose de résoudre des équations sans qu'il ait discuté la position physique du problème.

Naturelle? Que non! En effet, le mathématicien a souvent à poser un problème, à analyser et à simplifier les équations obtenues. Puisque tous les phénomènes de la nature sont en interaction, il est à priori impossible de mettre sous une forme mathématique rigoureuse un processus quelconque se déroulant réellement. Néanmoins, en analysant l'ascendant des divers facteurs sur l'erreur dans la solution, on obtiendrait une description élémentaire du processus avec une erreur admissible. En général, le mathématicien se fait une idée de la précision désirée et il effectue en conséquence les simplifications nécessaires.

Si le mathématicien ne discute pas la position physique du problème, il doit connaître l'ordre de grandeur de l'erreur inhérente. En effet, lorsqu'il résout un problème en se fondant sur le seul « bon sens », il n'a aucune raison particulière à appliquer une méthode qui donne la solution avec une erreur largement inférieure à l'erreur inhérente. Par conséquent, s'il connaît l'ordre de grandeur de l'erreur inhérente, le chercheur peut réduire les conditions imposées à la précision de l'algorithme qu'il applique.

On a eu un jour à résoudre le problème suivant: pour le plus grand nombre de domaines G et de deuxièmes membres f dresser les tableaux des solutions de l'équation de Poisson:

$$\Delta u = u_{x_1x_1} + u_{x_2x_2} = f(x_1, x_2) \quad \text{dans} \quad G,\ u|_\Gamma = 0;$$

Γ est la frontière de G, la précision exigée des valeurs tabulées devait être de l'ordre de 10^{-6}. Ainsi formulé, le problème était insoluble dans les délais admissibles, et on s'est demandé s'il avait été convenablement posé.

Il s'est avéré que le phénomène physique était régi par l'équation différentielle

$$(k(x_1, x_2)\, u_{x_1})_{x_1} + (k(x_1, x_2)\, u_{x_2})_{x_2} = f(x_1, x_2);$$

on ne savait rien du coefficient $k(x_1, x_2)$ sinon que pour les procédés de fabrication adoptés du matériau considéré il était compris entre 0,7 et 1,5. Il est

clair que la résolution numérique du problème initial avec une erreur inférieure à 0,01 (et peut-être à 0,1) était une absurdité. En diminuant la précision, on a pu réduire le nombre de domaines et de deuxièmes membres vu que les solutions dans des domaines voisins et pour des deuxièmes membres voisins sont voisines elles aussi. Désormais, le problème admettait une solution.

On voit donc qu'il est parfois commode d'étudier l'erreur inhérente à la position initiale du problème afin de réduire les conditions imposées à la précision requise.

Certains auteurs adoptent une autre classification d'erreurs. L'*erreur inhérente* est pour eux l'erreur commise sur les données, et l'approximation de la réalité par des modèles mathématiques est dite *erreur de formulation*.

§ 2. Enregistrement des nombres sur machines à calculer électroniques

Passons en revue deux formes de représentation des nombres les plus utilisées pour les ordinateurs modernes.

Dans l'écriture *en virgule fixe* tous les nombres sont inférieurs à 1 en module et le nombre de digits après la virgule est fixe. La machine opère donc avec des nombres

$$x = \pm \sum_{k=1}^{t} \alpha_k q^{-k} = \pm (\alpha_1, \ldots, \alpha_t); \qquad (1)$$

l'entier q est la base du système de numération, et $\alpha_1, \ldots, \alpha_t$ des entiers compris dans les limites $0 \leqslant \alpha_k < q$.

En travaillant sur des x tels que $|x| < 1$, l'apparition éventuelle de y, $|y| \geqslant 1$, entraîne l'arrêt de la machine. Pour l'éviter, on introduit une nouvelle échelle. Parfois il est difficile d'établir cette échelle à l'avance; dans d'autres cas, une échelle très grande fait dès le début s'annuler en quantité importante les premiers α_t des données initiales, ce qui signifie une perte d'information considérable. Le changement d'échelle s'effectue donc souvent en cours de résolution.

Avec la représentation *en virgule flottante*, la machine opère avec des nombres

$$x = \pm q^p \sum_{k=1}^{t} \alpha_k q^{-k} = \pm q^p (\alpha_1, \ldots, \alpha_t); \qquad (2)$$

l'ordre de grandeur de p satisfait à $|p| \leqslant p_0$.

Le système de numération le plus répandu est le code binaire ($q = 2$). Le décimal codé binaire ($q = 10$) et l'octal ($q = 8$) sont utilisés à l'entrée et à la sortie de la calculatrice. Le code ternaire ($q = 3$) est d'un emploi très restreint. Les utilisateurs des ordinateurs les plus évolués utilisent les deux types de représentation, et ils peuvent parfois choisir la taille des nombres.

Le travail en virgule flottante a l'avantage de tourner les difficultés liées à l'échelle des problèmes, cependant le fonctionnement de la machine s'en trouve ralenti.

Les valeurs les plus usuelles sont $q = 2$, $t = 36$, $p_0 = 64$. Alors $q^{-t} = 2^{-36} \approx 3 \cdot 10^{-11}$, $2^{p_0} = 2^{64} \approx 7 \cdot 10^{19}$.

§ 3. Erreur absolue et erreur relative. Ecritures des données

Si a est la valeur exacte d'une grandeur et a^* une valeur approchée, on appelle *erreur absolue* sur a^* et on note $\Delta(a^*)$ une quantité

$$|a^* - a| \leqslant \Delta(a^*).$$

On appelle *erreur relative* une quantité $\delta(a^*)$ telle que

$$\left|\frac{a^* - a}{a^*}\right| \leqslant \delta(a^*).$$

L'erreur relative est souvent exprimée sous forme d'un pourcentage.

Vous rencontrerez par la suite des expressions telles qu'un *nombre grand*, un *nombre très grand*, un *fort accroissement d'une fonction*.

On dit que x est *grand* si $|x| \gg 1$ et si l'erreur relative en $|x| 2^{-t}$ commise sur la solution du problème est permise;

si l'erreur relative de l'ordre de $|x| 2^{-t}$ est inacceptable, on dit que x est *très grand*;

par *fort accroissement d'une fonction* on entend que cette fonction croît de plusieurs fois.

On définit parfois: erreur absolue $a^* - a$; erreur relative $\frac{a^* - a}{a^*}$. Nous adopterons la première définition et supposerons toujours $0 \leqslant \Delta(a^*)$, $\delta(a^*)$.

On appelle *chiffres significatifs* d'un nombre tous les chiffres de son écriture à partir du premier chiffre différent de zéro à gauche.

EXEMPLE. Les chiffres significatifs des nombres $a^* = 0{,}03045$ et $a^* = 0{,}03045000$ sont ceux soulignés. Ils sont 4 dans le premier cas et 7 dans le deuxième.

Un chiffre significatif est dit *exact* si l'erreur absolue sur le nombre ne dépasse pas l'unité de l'ordre correspondant.

EXEMPLES. $a^* = 0{,}03045$, $\Delta(a^*) = 0{,}000003$; $a^* = 0{,}03045000$, $\Delta(a^*) = 0{,}0000007$; les chiffres soulignés sont exacts.

Si tous les chiffres significatifs sont exacts, on dit que le nombre est écrit *avec tous les chiffres exacts*.

EXEMPLE. Pour $a^* = 0{,}03045$, $\Delta(a^*) = 0{,}000003$, a^* est écrit avec 4 chiffres exacts.

On parle quelquefois de *décimales exactes*. Dans le dernier exemple le nombre a^* est écrit avec 5 décimales exactes.

Très souvent l'information sur une grandeur est fournie par l'intervalle de variation de celle-ci :

$$a_1 \leqslant a \leqslant a_2, \tag{1}$$

ainsi,

$$1{,}119 \leqslant a \leqslant 1{,}127.$$

Ces limites sont généralement écrites avec un même nombre de décimales. Puisque d'habitude on se contente d'une évaluation grossière de l'erreur, les limites a_1, a_2 sont prises de telle façon que leur différence $a_2 - a_1$ contienne de un à trois chiffres significatifs.

Ces réserves que nous faisons en disant « souvent », « d'habitude », « généralement » sont destinées à écarter l'idée d'une information donnée sous une forme standard obligatoire. Si ces formes nous intéressent, c'est qu'elles sont les plus usuelles, donc les plus commodes dans les communications.

Les erreurs absolues et relatives sont généralement représentées par des nombres à un ou deux chiffres significatifs.

L'information que a^* est une valeur approchée de a entachée de l'erreur absolue $\Delta(a^*)$ se met parfois sous forme commode

$$a = a^* \pm \Delta(a^*); \tag{2}$$

par convention on écrit a^* et $\Delta(a^*)$ avec un même nombre de décimales. Ainsi, les écritures

$$a = 1{,}123 \pm 0{,}004,$$
$$a = 1{,}123 \pm 4 \cdot 10^{-3}$$

sont parmi les plus usuelles et signifient que

$$1{,}123 - 0{,}004 \leqslant a \leqslant 1{,}123 + 0{,}004.$$

De même, l'information que a^* est une valeur approchée de a affectée de l'erreur relative $\delta(a^*)$ s'écrit

$$a = a^* (1 \pm \delta(a^*)). \tag{3}$$

Les écritures

$$a = 1{,}123 (1 \pm 0{,}003),$$
$$a = 1{,}123 (1 \pm 3 \cdot 10^{-3}),$$
$$a = 1{,}123 (1 \pm 0{,}3\ \%)$$

veulent dire

$$(1 - 0{,}003)\ 1{,}123 \leqslant a \leqslant (1 + 0{,}003)\ 1{,}123.$$

En passant d'une écriture à l'autre, il faut veiller à ce que les nouvelles limites de variation soient plus larges que les anciennes. Dans le cas contraire, ce passage n'est pas permis. Ainsi, en passant

de (1) à (2) on doit avoir

$$a^* - \Delta(a^*) \leqslant a_1, \quad a_2 \leqslant a^* + \Delta(a^*),$$

et en passant de (2) à (3),

$$a^*(1 - \delta(a^*)) \leqslant a^* - \Delta(a^*),\ a^* + \Delta(a^*) \leqslant a^*(1 + \delta(a^*)).$$

Le passage de (3) à (2) donne lieu aux inégalités de sens opposé (il y a toujours élargissement des limites!).

§ 4. Erreur de calcul

La condition $|p| \leqslant p_0$ conduit parfois à l'arrêt de la machine; il arrive également qu'une faible longueur des nombres traités donne lieu à une erreur de calcul inadmissible. Les algorithmes qui se prêtent à de tels effets à cause d'un p limité ou d'un t petit sont dits *instables*.

La théorie des méthodes numériques comprend comme partie importante l'élaboration d'algorithmes stables tels que les résultats de calcul ne soient pas trop altérés par une erreur.

Nous démontrerons plus loin qu'une meilleure précision peut quelquefois résulter d'une transformation algébrique simple.

On désire trouver la plus petite racine de l'équation $y^2 - 140y + 1 = 0$. Adoptons le système de numération décimal et supposons que la mantisse arrondie contient quatre chiffres. On a

$$y = 70 - \sqrt{4899}, \qquad \sqrt{4899} = 69{,}992\ldots;$$

et après l'arrondissement

$$\sqrt{4899} \approx 69{,}99, \quad y \approx 70 - 69{,}99 = 0{,}01.$$

La même valeur de y peut s'écrire, si l'on « chasse l'irrationnalité du numérateur », comme suit:

$$y = 1/(70 + \sqrt{4899}),\ \sqrt{4899} \approx 69{,}99,\ 70 + 69{,}99 = 139{,}99$$

et après l'arrondissement

$$70 + 69{,}99 \approx 140{,}0.$$

Enfin

$$1/140 = 0{,}00714285\ldots,$$

et après l'arrondissement

$$y \approx 0{,}007143.$$

Si le nombre de chiffres de la mantisse est supérieur à 4, on vérifie que dans les deux cas tous les chiffres soulignés sont exacts, le second résultat étant sensiblement plus exact. En effet, dans le premier cas, la soustraction a porté sur des nombres grands voisins.

Vu qu'il s'agit de nombres grands, ils ont été arrondis avec une erreur absolue importante. Nous nous heurtons ici au phénomène de perte de chiffres significatifs propre à la soustraction de quantités voisines. Ce phénomène altère souvent les solutions des systèmes d'équations algébriques.

§ 5. Erreur sur une fonction

On rencontre souvent le problème suivant. La quantité cherchée y est fonction des paramètres $a_1, \ldots, a_n$: $y = y(a_1, \ldots, a_n)$; le domaine G contenant ces paramètres est connu et contenu dans l'espace des variables $(a_1, \ldots, a_n)$. On demande d'obtenir une valeur approchée de y et d'évaluer l'erreur commise.

Soit y^* cette valeur approchée. On appelle alors *borne supérieure de l'erreur absolue* et on note $A(y^*)$ la meilleure évaluation, pour l'information dont on dispose, de l'erreur commise sur y^*; on a par définition

$$A(y^*) = \sup_{(a_1, \ldots, a_n) \in G} |y(a_1, \ldots, a_n) - y^*|; \tag{1}$$

la quantité $\frac{A(y^*)}{|y^*|}$ est la *borne supérieure de l'erreur relative.*

PROBLÈME. Démontrer que $A(y^*)$ est minimale pour

$$y^* = \frac{Y_1 + Y_2}{2},$$

avec

$$Y_1 = \inf_G y(a_1, \ldots, a_n), \qquad Y_2 = \sup_G y(a_1, \ldots, a_n).$$

Considérons le cas le plus répandu de G un rectangle:

$$|a_j - a_j^*| \leqslant \Delta(a_j^*), \quad j = 1, \ldots, n,$$

où l'on a pour valeur approchée

$$y^* = y(a_1^*, \ldots, a_n^*).$$

Si y est une fonction continûment dérivable par rapport à ses arguments, la formule de Lagrange donne

$$y(a_1, \ldots, a_n) - y^* = \sum_{j=1}^{n} b_j(\theta)(a_j - a_j^*),$$

où

$$b_j(\theta) = y_{a_j}(a_1^* + \theta(a_1 - a_1^*), \ldots, a_n^* + \theta(a_n - a_n^*)), \quad 0 \leqslant \theta \leqslant 1.$$

D'où l'estimation

$$|y(a_1, \ldots, a_n) - y^*| \leqslant A_0(y^*) = \sum_{j=1}^{n} B_j \Delta(a_j^*), \tag{3}$$

avec

$$B_j = \sup_G | y_{a_j}(a_1, \ldots, a_n) |.$$

Montrons que pour

$$\rho = \sqrt{(\Delta(a_1^*))^2 + \ldots + (\Delta(a_n^*))^2}$$

petits, cette estimation ne se prête pas à une amélioration sensible. Si les dérivées y_{a_j} sont continues dans G, alors

$$B_j = | b_j(0) | + o(1) \text{ pour } \rho \to 0.$$

Et $A_0(y^*)$ peut s'écrire

$$\begin{gathered} A_0(y^*) = A^0(y^*) + \varepsilon_1(\rho), \\ A^0(y^*) = \sum_{j=1}^{n} | b_j(0) | \Delta(a_j^*), \quad \varepsilon_1(\rho) = o(\rho). \end{gathered} \tag{4}$$

Pour

$$a_j - a_j^* = \operatorname{sg}(b_j(0))\, \Delta(a_j^*)$$

on a

$$\begin{aligned} b_j(\theta)\,(a_j - a_j^*) &= (b_j(0) + o(1))\operatorname{sg}(b_j(0))\,\Delta(a_j^*) = \\ &= (| b_j(0) | + o(1))\,\Delta(a_j^*), \end{aligned}$$

et alors

$$y(a_1, \ldots, a_n) - y^* = A^0(y^*) + o(\rho). \tag{5}$$

Il en découle, compte tenu de la définition (1) de la borne supérieure de l'erreur absolue, que

$$A^0(y^*) + \varepsilon_2(\rho) \leqslant A(y^*), \quad \varepsilon_2(\rho) = o(\rho), \tag{6}$$

et (3) entraîne

$$A(y^*) \leqslant A_0(y^*).$$

En réunissant ces deux expressions, il vient

$$A^0(y^*) + \varepsilon_2(\rho) \leqslant A(y^*) \leqslant A_0(y^*) = A^0(y^*) + \varepsilon_1(\rho). \tag{7}$$

Ainsi, l'estimation (3) n'est pas de beaucoup supérieure à la valeur exacte (1). Selon (7), $A(y^*)$ et $A_0(y^*)$ sont approchées avec une précision élevée par la quantité $A^0(y^*)$ qui se calcule plus simplement.

L'estimation (3) est souvent remplacée par

$$| y(a_1, \ldots, a_n) - y^* | \leqslant A^0(y^*), \tag{8}$$

qui n'est pas en fait exacte et qui s'appelle *estimation linéaire*.

Examinons quelques exemples de définition de $A(y^*)$, $A_0(y^*)$ et $A^0(y^*)$ et comparons ces grandeurs.

1. $y = a^{10}$, $a^* = 1$, $\Delta(a^*) = 0{,}001$. Alors

$$y^* = 1, \qquad y_a = 10 \cdot a^9, \quad b(0) = 10,$$
$$B = \sup_{|a-1| \leqslant 0{,}001} |10 \cdot a^9| = 10{,}09 \ldots,$$
$$A(y^*) = \sup_{|a-1| \leqslant 0{,}001} |a^{10} - 1| = 1{,}001^{10} - 1 = 0{,}010045 \ldots,$$
$$A_0(y^*) = B\Delta(a^*) = 0{,}01009 \ldots,$$
$$A^0(y^*) = |b(0)| \Delta(a^*) = 0{,}01.$$

Ici, les trois évaluations diffèrent peu.

2. $y = a^{10}$, $a^* = 1$, $\Delta(a^*) = 0{,}1$. Alors

$$y^* = 1, \quad B = \sup_{|a-1| \leqslant 0{,}1} 10 \cdot a^9 = 10 \cdot (1{,}1)^9 = 23, \ldots,$$
$$A(y^*) = \sup_{|a-1| \leqslant 0{,}1} |a^{10} - 1| = (1{,}1)^{10} - 1 = 1{,}5 \ldots,$$
$$A_0(y^*) = B\Delta(a^*) = 2{,}3 \ldots, \quad A^0(y^*) = |b(0)| \Delta(a^*) = 1.$$

La différence s'avère plus prononcée.

3. On se propose maintenant d'évaluer des fonctions élémentaires. Soit

$$y = \gamma_1 a_1 + \ldots + \gamma_n a_n,$$

$\gamma_1, \ldots, \gamma_n$ prenant les valeurs $+1$ ou -1; supposons qu'on connaît les estimations $|a_j - a_j^*| \leqslant \Delta(a_j^*)$. Dans notre cas $b_j(\theta) \equiv \gamma_j$; $|b_j(\theta)| \equiv 1$, aussi les termes $\varepsilon_j(\rho)$ d'ordre supérieur sont absents dans (4), (6), et (7) implique

$$A(y^*) = A_0(y^*) = A^0(y^*) = \Delta(a_1^*) + \ldots + \Delta(a_n^*).$$

Puisque, par définition, on appelle erreur toute estimation de $y - y^*$, cette relation s'écrit de même

$$\Delta(\pm a_1^* \pm \ldots \pm a_n^*) = \Delta(a_1^*) + \ldots + \Delta(a_n^*). \tag{9}$$

Si $y = a_1 \pm a_2$, cette égalité s'énonce sous forme de règle:

la borne supérieure de l'erreur absolue sur une somme ou une différence est égale à la somme des bornes supérieures des erreurs absolues sur les termes.

Si les erreurs commises sur a_j^* sont liées entre elles, l'estimation (9) peut en général être améliorée. En voici un exemple élémentaire: $a_1 = a$, $a_2 = 1 - a$, avec un a identique dans les deux cas; alors quelle que soit l'erreur sur a, la somme $a_1 + a_2$ est égale à 1.

Soit maintenant $y = a_1^{p_1} \ldots a_n^{p_n}$; alors $b_j(0) = p_j (a_j^*)^{-1} y^*$ et

$$A(y^*) \approx A^0(y^*) = \sum_{j=1}^{n} |p_j| |a_j^*|^{-1} |y^*| \Delta(a_j^*).$$

En divisant par $|y^*|$ on obtient

$$\frac{A(y^*)}{|y^*|} \approx \frac{A^0(y^*)}{|y^*|} = \sum_{j=1}^{n} |p_j| \frac{\Delta(a_j^*)}{|a_j^*|} = \sum_{j=1}^{n} |p_j| \delta(a_j^*). \tag{10}$$

Pour les cas particuliers $y = a_1 \cdot a_2$ ou $y = a_1 \cdot a_2^{-1}$, cette relation s'énonce quelquefois comme une règle :

la borne supérieure de l'erreur relative sur un produit (resp. *un quotient*) *est approximativement égale à la somme des bornes supérieures des erreurs relatives sur les facteurs* (resp. *le dividende et le diviseur*).

4. Un problème assez fréquent est l'évaluation de l'erreur commise sur une fonction définie implicitement par l'équation

$$F(y, a_1, \ldots, a_n) = 0. \tag{11}$$

Dérivant par rapport à a_j, il vient

$$\frac{\partial F}{\partial y} \frac{\partial y}{\partial a_j} + \frac{\partial F}{\partial a_j} = 0,$$

d'où

$$\frac{\partial y}{\partial a_j} = -\left(\frac{\partial F}{\partial a_j}\right)\left(\frac{\partial F}{\partial y}\right)^{-1}. \tag{12}$$

Quand $a_1^*, \ldots, a_n^*$ sont donnés, on trouve y^* comme racine de (11), puis

$$b_j(0) = -\left(\frac{\partial F}{\partial a_j}\right)\left(\frac{\partial F}{\partial y}\right)^{-1}\Big|_{(y^*, a_1^*, \ldots, a_n^*)}. \tag{13}$$

En faisant appel à ces quantités, on obtient l'« estimation linéaire » d'erreur (8).

Les dérivées $\frac{\partial y}{\partial a_j}$ étant fonction de y, le calcul des estimations strictes (1), (3) s'avère assez coûteux.

5. Evaluons l'erreur sur les racines de l'équation du second degré

$$F(y, a_1, a_2) = y^2 + a_1 y + a_2 = 0 \tag{14}$$

connaissant les valeurs approchées des coefficients a_1^*, a_2^* et leurs erreurs $\Delta(a_1^*)$, $\Delta(a_2^*)$.

Soit y^* une solution de

$$y^{*2} + a_1^* y^* + a_2^* = 0.$$

La formule (13) fournit

$$b_1(0) = \frac{\partial y}{\partial a_1}\Big|_{(a_1^*, a_2^*)} = -\frac{y^*}{2y^* + a_1^*},$$

$$b_2(0) = \frac{\partial y}{\partial a_2}\Big|_{(a_1^*, a_2^*)} = -\frac{1}{2y^* + a_1^*},$$

donc

$$A^0(y^*) = \frac{|y^*|\,\Delta(a_1^*) + \Delta(a_2^*)}{|2y^* + a_1^*|}. \qquad (15)$$

Considérons un domaine $|a_1| \leqslant b_1$, $|a_2| \leqslant b_2$ de variation des coefficients a_1, a_2. L'expression explicite des racines

$$y = -\frac{a_1}{2} \pm \sqrt{\frac{a_1^2}{4} - a_2}$$

entraîne que les racines sont des fonctions continues des coefficients, aussi

$$|y(a_1, a_2) - y(a_1^*, a_2^*)| \leqslant \omega(|a_1 - a_1^*|, |a_2 - a_2^*|) \qquad (16)$$

pour (a_1, a_2), (a_1^*, a_2^*) du domaine, $\omega(\lambda_1, \lambda_2) \to 0$ lorsque $\lambda_1, \lambda_2 \to 0$. Le second membre de (15) tend vers l'infini quand $2y^* + a_2^* \to 0$; en comparant (15) et (16) on voit donc que l'« estimation linéaire » d'erreur par la formule (8) peut être trop forte par rapport à l'estimation exacte (3). En effet, dans nos raisonnements nous avons supposé $y(a_1, \ldots, a_n)$ continûment dérivable par rapport aux arguments $a_1, \ldots, a_n$, et l'erreur sur y^* était du même ordre de grandeur que $\Delta(a_j^*)$, erreur commise sur les arguments. Si y^* est définie sous une forme implicite, elle n'est pas dérivable par rapport à ses arguments pour certaines valeurs des paramètres, et l'évaluation change en conséquence.

Soit y^* une racine double de (11) pour $a_1 = a_1^*$, $a_2 = a_2^*$. Développons le premier membre de cette équation en série de Taylor dans le voisinage du point y^*, a_1^*, a_2^*. Puisque

$$F(y^*, a_1^*, a_2^*) = F_y(y^*, a_1^*, a_2^*) = 0$$

quand y^* est racine double de (11), celle-ci s'écrit

$$d_{200}(y - y^*)^2 + d_{010}(a_1 - a_1^*) + d_{001}(a_2 - a_2^*) + \ldots = 0,$$

avec

$$d_{ijk} = \frac{F_{y^i a_1^j a_2^k}(y^*, a_1^*, a_2^*)}{i!\,j!\,k!},$$

les termes négligés étant de l'ordre de $o(\rho)$. Dans le cas (14) $d_{200} = 1$ et alors

$$y - y^* = \pm\sqrt{d_{010}(a_1 - a_1^*) + d_{001}(a_2 - a_2^*) + o(\rho)}.$$

Donc, l'erreur qui entache la racine est une quantité en $O(\sqrt{\rho})$. Partant, l'erreur sur une racine d'ordre de multiplicité l est d'ordinaire de l'ordre de $O(\sqrt[l]{\rho})$.

6. L'évaluation d'erreurs sur les racines d'équations algébriques peut poser un problème qui de prime abord semble opportun et natu-

rel. On considère l'équation

$$y^n + \sum_{j=1}^{n} a_j^* y^{n-j} = 0. \tag{17}$$

Soient donnés un domaine de variation des valeurs des coefficients a_j^* et l'estimation $\Delta(a_j^*) \leqslant \gamma$. Estimer la borne supérieure ω des erreurs sur les racines de cette équation.

PROBLÈME. Etablir la même évaluation pour l'équation (17) avec $|a_1^*|$, $|a_2^*| \leqslant 2$, $\Delta(a_1^*)$, $\Delta(a_2^*) \leqslant \gamma$.

Si le domaine de variation des a_j^* est tel que pour tout élément $\{a_j^*\}$ de ce domaine l'équation peut posséder une racine l-uple, on ne peut espérer une évaluation d'un ordre meilleur que $O(\sqrt[l]{\gamma})$. Cependant, la racine de l'équation est simple et l'ordre de grandeur $O(\gamma)$ est possible à l'exception d'un ensemble de valeurs des coefficients de mesure nulle. Ainsi, une estimation d'erreur de l'ordre de $O(\sqrt[l]{\gamma})$ serait trop forte pour la plupart des $a_1^*, \ldots, a_n^*$, ce qui met en doute son utilité.

7. Voici un autre problème type fréquent dans les applications et qui se résout d'une façon semblable.

Soit y^* une approximation d'une racine de l'équation

$$f(y) = a;$$

on demande d'évaluer l'erreur commise.

Calculons

$$a^{*\prime} = f(y^*).$$

Pour $y^* - y$ petits, l'égalité

$$f(y) - f(y^*) = a - a^*$$

entraîne

$$f'(y^*)(y - y^*) \approx a - a^*,$$

donc

$$y - y^* \approx \frac{a - a^*}{f'(y^*)} = \frac{a - f(y^*)}{f'(y^*)}.$$

Pour $a = 0$, cas le plus répandu, il vient

$$y - y^* \approx -\frac{f(y^*)}{f'(y^*)}.$$

CHAPITRE II

INTERPOLATION ET PROBLÈMES CONNEXES

L'approximation des fonctions est un problème qui est susceptible de se résoudre aussi bien séparément que dans le cadre d'autres problèmes plus vastes. Le présent chapitre est consacré à ce procédé d'approximation très répandu qu'est l'interpolation. L'interpolation est un outil auxiliaire de premier plan de l'analyse numérique, notamment dans l'intégration et la dérivation numériques, la résolution d'équations différentielles, etc. Nous voudrions au préalable rappeler au lecteur certaines définitions et discuter diverses façons de poser un problème d'approximation des fonctions (§ 1).

Un ensemble d'éléments M constitue un espace *vectoriel* s'il est muni des opérations internes d'addition et de multiplication par un nombre (réel ou complexe) satisfaisant aux conditions:

1) l'addition est associative: $(x + y) + z = x + (y + z)$,
2) l'addition est commutative: $x + y = y + x$,
3) il existe un élément nul 0 tel que $x + 0 = x$ quel que soit $x \in M$,
4) $0 \cdot x = 0$ pour tout $x \in M$,
5) $(\alpha + \beta)\, x = \alpha x + \beta x$,
6) $\alpha\,(x + y) = \alpha x + \alpha y$,
7) $\alpha\,(\beta x) = (\alpha\beta)\, x$,
8) $1 \cdot x = x$.

On introduit dans l'espace vectoriel la notion de dépendance et d'indépendance linéaires d'éléments. Un système d'éléments $x_1, \ldots, x_n$ d'un espace vectoriel M est dit *linéairement dépendant* s'il existe des $c_1, \ldots, c_n$ non tous nuls tels que

$$c_1 x_1 + \ldots + c_n x_n = 0.$$

Dans le cas contraire, le système est *linéairement indépendant*.

Un sous-ensemble H d'un espace vectoriel est un *sous-espace vectoriel* si $x, y \in H$ implique $\alpha x + \beta y \in H$ quels que soient α et β.

Un espace $\mathcal{R}$ est dit *métrique* si l'on peut faire correspondre à deux quelconques de ses éléments la distance $\rho(x, y)$ vérifiant les conditions:

1) $\rho(x, y) \geqslant 0$, $\rho(x, y) = 0$ si et seulement si $x = y$,
2) $\rho(x, y) = \rho(y, x)$,
3) $\rho(x, y) \leqslant \rho(x, z) + \rho(z, y)$ quels que soient $x, y, z \in \mathcal{R}$.

L'espace $\mathcal{R}$ s'appelle *espace vectoriel normé* si a) il est vectoriel et b) on peut faire correspondre à tout élément $f \in \mathcal{R}$ un nombre réel $\| f \|$ appelé *norme de* f, tel que

1) $\| f \| \geqslant 0$, $\| f \| = 0$ si et seulement si $f = 0$,
2) $\| \alpha f \| = | \alpha | \, \| f \|$ pour tout α complexe,
3) $\| f_1 + f_2 \| \leqslant \| f_1 \| + \| f_2 \|$.

Il est évident qu'un espace vectoriel normé est un espace métrique avec pour distance

$$\rho(f_1, f_2) = \| f_1 - f_2 \|.$$

On dit d'un espace vectoriel normé qu'il est *strictement normé* si on a

$$\| f_1 + f_2 \| = \| f_1 \| + \| f_2 \|$$

si et seulement si $f_2 = \alpha f_1$, $\alpha \geqslant 0$.

On dit que le produit scalaire est défini dans l'espace vectoriel $\mathcal{R}$ si à tout couple ordonné d'éléments $f_1, f_2 \in \mathcal{R}$ on a fait correspondre un nombre complexe (f_1, f_2) et si

1. $(f_1, f_2) = \overline{(f_2, f_1)}$.
2. Quels que soient $f_1, f_2, f_3 \in \mathcal{R}$ et α_1, α_2 complexes, on a

$$(\alpha_1 f_1 + \alpha_2 f_2, f_3) = \alpha_1 (f_1, f_3) + \alpha_2 (f_2, f_3).$$

3. $(f, f) \geqslant 0$ et $(f, f) = 0$ seulement pour $f = 0$.

Ces propriétés du produit scalaire en entraînent d'autres, notamment:

4. $(f_2, \alpha_1 f_1 + \alpha_2 f_2) = \overline{\alpha}_1 (f_2, f_1) + \overline{\alpha}_2 (f_2, f_2)$.

Quels que soient $f_1, f_2 \in \mathcal{R}$, on a les inégalités:

5. $|(f_1, f_2)|^2 \leqslant (f_1, f_1)(f_2, f_2)$.
6. $\| f_1 + f_2 \| \leqslant \| f_1 \| + \| f_2 \|$, où $\| f \| = \sqrt{(f, f)}$.
7. Le signe d'égalité de la propriété 6 n'a lieu que pour $f_2 = \alpha f_1$, $\alpha \geqslant 0$.

La propriété 6 signifie que l'espace vectoriel $\mathcal{R}$ muni du produit scalaire est un espace vectoriel normé, et, par conséquent, un espace métrique avec

$$\rho(f_1, f_2) = \| f_1 - f_2 \| = \sqrt{(f_1 - f_2, f_1 - f_2)}.$$

La propriété 7 traduit le fait que c'est un espace strictement normé.

§ 1. Approximation des fonctions. Position du problème

Le contexte du problème d'approximation d'une fonction peut être, on le verra dans la suite, des plus divers. Les techniques proposées sont si nombreuses qu'on se demande parfois si cette diversité n'est pas due à l'absence d'une approche scientifique. Si cette approche avait existé, peut-être aurait-elle permis de dégager une méthode d'approximation à la fois optimale et universelle. On se pose d'ailleurs la même question en étudiant les autres branches de l'analyse numérique. Pourtant, quelque captivante que paraisse l'idée d'une approche universelle, la variété des méthodes est une conséquence logique de l'abondance des façons dont on pose le problème. Les différents chapitres de la théorie des approximations, et notamment l'interpolation, peuvent être assimilés à l'étude des modèles abstraits de certaines classes concrètes de problèmes.

1. Ci-dessous un problème simple conduisant à l'approximation des fonctions. On observe aux instants discrets $x_1, \ldots, x_n$ les valeurs d'une fonction $f(x)$; on demande de restituer ses valeurs pour d'autres x. Le même problème se pose dans les conditions suivantes. En effectuant des calculs sur machine l'on a à déterminer la valeur d'une fonction compliquée en divers points et à plusieurs reprises. Le calcul direct est alors parfois à remplacer par le calcul

des valeurs en des points isolés, et d'autres valeurs s'obtiennent moyennant cette information.

Il arrive qu'on sache, certaines considérations supplémentaires aidant, que la fonction d'approximation doit être de la forme

$$f(x) \approx g(x; a_1, \ldots, a_n).$$

Si les paramètres $a_1, \ldots, a_n$ sont définis par la condition de coïncidence de $f(x)$ et de fonction approximante aux points $x_1, \ldots, x_n$:

$$g(x_j; a_1, \ldots, a_n) = f(x_j), \quad j = 1, \ldots, n,$$

on est en présence d'une méthode d'approximation appelée *interpolation*. Si le point x n'appartient pas au segment $[\min x_i, \max x_i]$, on parle de l'extrapolation.

2. On sait souvent que la fonction est bien approchée par des fonctions d'une certaine forme, par exemple par des polynômes, mais on ignore la façon de choisir au mieux le degré du polynôme approximant. Le problème se complique singulièrement si les valeurs données de la fonction sont entachées de grosses erreurs.

3. En rédigeant un programme d'expériences, on se heurte au problème suivant. On connaît la forme d'un bon approximant de la fonction, mettons que ce soit un polynôme du deuxième degré. D'autre part, les valeurs de la fonction sont mesurées avec une grande erreur. On demande d'obtenir une meilleure approximation relativement à une certaine norme avec un nombre minimal de mesures. Ce problème se rencontre en biologie, chimie, physique, géographie et art militaire.

Parlerons plus en détail des plans d'expériences dans l'industrie chimique. Le paramètre de sortie essentiel du produit, par exemple la qualité du Kapron, est défini par une série de paramètres dont certains sont fonctions de la matière première et d'autres sont contrôlables. On demande d'augmenter au maximum la production de Kapron de première qualité. Pour le faire on doit former une fonction traduisant les relations qui existent entre la productivité et ces paramètres. Considérons le cas où cette fonction est un polynôme du deuxième degré en 10 paramètres. Ce polynôme contient 66 coefficients dont la recherche exige au moins 66 mesures. Chacune d'elles est entachée d'une erreur accidentelle essentielle due à des facteurs échappant à tout contrôle et à toute mesure. Aussi doit-on, pour réduire le rôle du hasard, augmenter le nombre de mesures. Par ailleurs, chaque mesure coûte relativement cher du moment qu'elle nécessite une modification du régime de travail de l'entreprise. Cette modification s'échelonne de plus sur une période prolongée parce que l'unité de production en question doit en fin de compte faire sien un régime de travail stationnaire satisfaisant aux valeurs données des paramètres. Le problème d'approximation des fonctions par lequel nous avons commencé le présent numéro est bien l'aboutissement de l'analyse de ces phénomènes contradictoires.

4. Un problème d'approximation se pose de même lorsqu'on élabore des programmes standard de calcul des fonctions élémentaires et spéciales. Ces fonctions possèdent d'habitude des propriétés particulières permettant de réduire notablement le volume de calcul.

Ce problème s'énonce comme suit. On considère toutes les fonctions $g(x)$ dont le programme de calcul est enregistré en k cellules de mémoire de la machine et telles qu'une certaine norme $\|f - g\|$ de l'erreur ne dépasse pas ε. Parmi ces fonctions choisir celle dont le calcul exige un temps machine minimal.

La norme est choisie selon les circonstances. Dans la plupart des cas on prend $\|f\| = \sup_{[a, b]} |f|$, $[a, b]$ étant le segment sur lequel on approche la fonction.

Il arrive souvent qu'on ait besoin d'une précision plus élevée en certains points. Ainsi, un programme usuel de calcul de $\sin x$ assure la petitesse de l'erreur en norme

$$\|f\| = \sup_{[0, \frac{\pi}{2}]} |p(x) f(x)|, \quad p(x) = \min(10^{19}, x^{-1}).$$

On a introduit le facteur $p(x)$ pour vérifier la condition de petitesse de l'erreur relative sur $\sin x$ pour x petits.

Il en est de même de la condition de temps machine minimal. Le coût peut en général dépendre du point en lequel on calcule la fonction. Soit $t(x)$ ce coût. Si l'on ne dispose pas d'information sur la fréquence avec laquelle on calcule la fonction sur telles ou telles parties du segment, on mesure les coûts par

$$T = \sup_x t(x).$$

En présence d'une telle information on a

$$T = \mathrm{E}t(x).$$

5. Si la fonction est donnée par sa courbe représentative ou par une expression analytique compliquée, on voit apparaître des problèmes variationnels de types différents. Convenons par exemple de partager le segment $[a, b]$ en l parties

$$[a_{i-1}, a_i], \quad i = 1, \ldots, l, \quad a_0 = a, \quad a_l = b,$$

puis d'approcher la fonction sur chaque segment partiel $[a_{i-1}, a_i]$ par le polynôme

$$f(x) \approx g(x) = \sum_{k=0}^{n_i} a_k^i x^k$$

ou la fonction rationnelle

$$f(x) \approx g(x) = \frac{\sum_{k=0}^{n_i} a_k^i x^k}{\sum_{k=0}^{m_i} b_k^i x^k}.$$

Parmi de telles méthodes d'approximation, on recherche celle qui est optimale dans un sens ou dans l'autre. Le procédé le plus usité est le suivant: on impose d'avance la condition $n_1 = \ldots = n_l$, $m_1 \ldots = m_l$, on fixe le nombre l de segments partiels et on optimise la méthode pour $a_1, \ldots, a_{l-1}$, n et m.

Dans le cas particulier $l = 1$, on a le problème de meilleures approximations uniformes par des polynômes ou des fonctions rationnelles sur un segment fixe. On en reparlera au chapitre IV.

6. La forme de la fonction qui approche dépend essentiellement du but de l'approximation. Supposons que la fonction est approchée avec le degré de précision voulu par un polynôme du dixième degré ou par l'expression $a_1 \sin \omega_1 x + a_2 \sin \omega_2 x$. La deuxième écriture est plus commode à utiliser dans un problème théorique, la simulation ou dans un processus de fabrication; elle nécessite par contre bien plus d'opérations arithmétiques dans le cas où la fonction se calcule sur machine.

Dans la suite nous aborderons l'interpolation dans le détail. Si nous avons mis l'accent sur ce problème, c'est qu'il se connaît de nombreuses applications directes et aussi pour la raison suivante. L'interpolation polynomiale, outil majeur de l'analyse numérique, est à la base de la plupart des méthodes de résolution d'autres problèmes. Elle joue en analyse numérique le même rôle que le développement taylorien en analyse classique.

On étudiera parallèlement certaines questions générales dont l'importance pour d'autres chapitres de l'analyse numérique est certaine.

§ 2. Polynôme d'interpolation de Lagrange

De toutes les techniques d'interpolation la plus utilisée est l'interpolation linéaire, l'approximation étant cherchée sous la forme

$$g(x;\ a_1,\ \ldots,\ a_n) = \sum_{i=1}^{n} a_i \varphi_i(x),$$

où $\varphi_i(x)$ sont fixes et les coefficients a_i sont déterminés par la condition de coïncidence avec la fonction approchée aux *points d'interpolation* x_j:

$$f(x_j) = \sum_{i=1}^{n} a_i \varphi_i(x_j), \quad j = 1,\ \ldots,\ n. \tag{1}$$

Le procédé permettant de déterminer a_i en résolvant directement le système (1) est connu sous le nom de *méthode des coefficients indéterminés.*

Le cas le plus étudié est celui de l'interpolation par les polynômes

$$\sum_{i=1}^{n} a_i x^{i-1}. \tag{2}$$

Alors

$$\varphi_i(x) = x^{i-1}, \quad i = 1, \ldots, n,$$

et le système (1) devient

$$\sum_{i=1}^{n} a_i x_j^{i-1} = f(x_j), \quad j = 1, \ldots, n. \tag{3}$$

Nous supposons implicitement que les x_j sont distincts. Le déterminant du système, à savoir $\det[x_j^{i-1}]$, étant différent de zéro (déterminant de Vandermonde), le système (3) admet toujours une solution et une seule. On a donc démontré l'existence et l'unicité du polynôme d'interpolation de la forme (2).

Il nous sera possible d'expliciter le polynôme (2) sans faire appel à la résolution directe du système (3). Notons néanmoins que dans d'autres cas on n'y arrive qu'avec beaucoup de difficultés de sorte qu'on est obligé de résoudre directement le système (1).

Soit $\delta_{q_1, \ldots, q_k}$ un symbole tel que

$$\delta_{q_1, \ldots, q_k} = \begin{cases} 1 \text{ pour } q_1 = \ldots = q_k, \\ 0 \text{ pour les autres jeux de valeurs } q_1, \ldots, q_k. \end{cases}$$

Le problème d'interpolation est résolu si l'on construit des polynômes $\Phi_i(x)$ de degré $n-1$ tels que $\Phi_i(x_j) = \delta_{ij}$ pour $i, j = 1, \ldots, n$. Le polynôme

$$g_n(x) = \sum_{j=1}^{n} f(x_j)\, \Phi_j(x)$$

est le polynôme d'interpolation cherché; en effet,

$$g_n(x_i) = \sum_{j=1}^{n} f(x_j)\, \Phi_j(x_i) = \sum_{j=1}^{n} f(x_j)\, \delta_{ij} = f(x_i);$$

de plus, le degré de $g_n(x)$ est $n-1$.

Vu que $\Phi_i(x_j) = 0$ pour $j \neq i$, $\Phi_i(x)$ est divisible par $x - x_j$ lorsque $j \neq i$; nous connaissons donc $n-1$ diviseurs du polynôme de degré $n-1$, d'où

$$\Phi_i(x) = \text{const} \cdot \prod_{j \neq i} (x - x_j).$$

La condition $\Phi_i(x_i) = 1$ donne

$$\Phi_i(x) = \prod_{j \neq i} \frac{x - x_j}{x_i - x_j}.$$

Le polynôme d'interpolation (2) mis sous la forme

$$g_n(x) = L_n(x) = \sum_{i=1}^{n} f(x_i) \prod_{j \neq i} \frac{x - x_j}{x_i - x_j} \tag{4}$$

est appelé *polynôme d'interpolation de Lagrange.*

Introduisons la fonction

$$\omega_n(x) = \prod_{j=1}^{n} (x - x_j).$$

Il est immédiat que

$$\omega_n'(x) = \sum_{k=1}^{n} \left[\prod_{j \neq k} (x - x_j) \right];$$

pour $x = x_i$, $k \neq i$, le crochet s'annule parce que contenant le facteur $(x_i - x_i)$. Ainsi

$$\omega_n'(x_i) = \prod_{j \neq i} (x_i - x_j),$$

le coefficient $\prod_{j \neq i} \frac{x - x_j}{x_i - x_j}$ s'écrit

$$\frac{\omega_n(x)}{\omega_n'(x_i)(x - x_i)},$$

et le polynôme de Lagrange

$$L_n(x) = \sum_{i=1}^{n} \frac{f(x_i)\,\omega_n(x)}{\omega_n'(x_i)(x - x_i)}. \tag{5}$$

§ 3. Evaluation du reste du polynôme d'interpolation de Lagrange

Supposons $f^{(n)}(x)$ continue et évaluons la différence entre $f(x)$ et le polynôme d'interpolation $g_n(x)$ qu'on vient de construire. Posons

$$\varphi(z) = f(z) - g_n(z) - K\omega_n(z),$$

K étant choisi à partir de la condition $\varphi(x) = 0$, où x est le point en lequel on évalue l'erreur. On a visiblement

$$K = \frac{f(x) - g_n(x)}{\omega_n(x)}.$$

Pour un tel K, la fonction $\varphi(z)$ s'annule en $n+1$ points $x_1, \ldots, x_n, x$. En vertu du théorème de Rolle, la dérivée $\varphi'(z)$ s'annule en n points au moins. En appliquant ce théorème à $\varphi'(z)$, on obtient que sa dérivée $\varphi''(z)$ s'annule en $n-1$ points au moins. En poursuivant on aboutit au résultat suivant: $\varphi^{(n)}(z)$ s'annule au moins en un point ζ qui appartient au segment $[y_1, y_2]$:

$$y_1 = \min(x_1, \ldots, x_n, x), \quad y_2 = \max(x_1, \ldots, x_n, x).$$

Puisque

$$\varphi^{(n)}(z) = f^{(n)}(z) - Kn!,$$

la condition $\varphi^{(n)}(\zeta) = 0$ entraîne

$$K = \frac{f^{(n)}(\zeta)}{n!}.$$

La relation $\varphi(x) = 0$ s'écrit encore

$$f(x) - g_n(x) = \frac{f^{(n)}(\zeta)\,\omega_n(x)}{n!}, \quad \zeta \in [y_1, y_2], \tag{1}$$

expression du reste.

§ 4. Différences divisées et leurs propriétés

Le polynôme d'interpolation peut s'interpréter comme la généralisation d'une tranche de la série de Taylor.

La notion de *différence divisée* généralise celle de dérivée. Les différences divisées d'ordre zéro $f(x_i)$ coïncident avec les valeurs de la fonction $f(x_i)$; les différences divisées premières (du premier ordre) sont définies par l'égalité

$$f(x_i; x_j) = \frac{f(x_j) - f(x_i)}{x_j - x_i}, \tag{1}$$

les différences divisées secondes par

$$f(x_i; x_j; x_k) = \frac{f(x_j; x_k) - f(x_i; x_j)}{x_k - x_i}.$$

D'une manière générale, on définit les différences divisées d'ordre k $f(x_1; \ldots; x_{k+1})$ au moyen des différences d'ordre $k-1$ par la formule

$$f(x_1; \ldots; x_{k+1}) = \frac{f(x_2; \ldots; x_{k+1}) - f(x_1; \ldots; x_k)}{x_{k+1} - x_1}. \tag{2}$$

L'écriture $f(x_1; \ldots; x_k)$ est parfois remplacée par $(f)(x_1; \ldots; x_k)$ ou $[x_1; \ldots; x_k]$.

LEMME.

$$f(x_1; \ldots; x_k) = \sum_{j=1}^{k} \frac{f(x_j)}{\prod\limits_{i \neq j} (x_j - x_i)}. \tag{3}$$

Démontrons-le par récurrence; pour $k = 1$, (3) devient $f(x_1) = f(x_1)$; pour $k = 2$, (3) coïncide avec (1). Supposons que cette égalité est satisfaite pour $k \leqslant l$. Alors

$$f(x_1; \ldots; x_{l+1}) = \frac{f(x_2; \ldots; x_{l+1}) - f(x_1; \ldots; x_l)}{x_{l+1} - x_1} =$$

$$= \frac{1}{x_{l+1} - x_1} \left(\sum_{j=2}^{l+1} \frac{f(x_j)}{\prod\limits_{\substack{i \neq j \\ 2 \leqslant i \leqslant l+1}} (x_j - x_i)} - \sum_{j=1}^{l} \frac{f(x_j)}{\prod\limits_{\substack{i \neq j \\ 1 \leqslant i \leqslant l}} (x_j - x_i)} \right).$$

Quand $j \neq 1$, $l + 1$, le coefficient de $f(x_j)$ du dernier membre est

$$\frac{\dfrac{1}{\prod\limits_{\substack{i \neq j \\ 2 \leqslant i \leqslant l+1}} (x_j - x_i)} - \dfrac{1}{\prod\limits_{\substack{i \neq j \\ 1 \leqslant i \leqslant l}} (x_j - x_i)}}{x_{l+1} - x_1} =$$

$$= \frac{(x_j - x_1) - (x_j - x_{l+1})}{(x_{l+1} - x_1) \prod\limits_{\substack{i \neq j \\ 1 \leqslant i \leqslant l+1}} (x_j - x_i)} = \frac{1}{\prod\limits_{\substack{i \neq j \\ 1 \leqslant i \leqslant l+1}} (x_j - x_i)},$$

i. e. il est de la forme demandée; pour $j = 1$ ou $j = l + 1$, $f(x_j)$ ne figure que dans un terme du dernier membre et son coefficient est lui aussi de la forme demandée. Le lemme est démontré.

La relation (3) donne lieu aux conséquences immédiates:

1. Une différence divisée est un opérateur linéaire de la fonction f:

$$(\alpha_1 f_1 + \alpha_2 f_2)(x_1; \ldots; x_k) = \alpha_1 f_1(x_1; \ldots; x_k) + \alpha_2 f_2(x_1; \ldots; x_k).$$

2. Une différence divisée est une fonction symétrique par rapport à ses arguments $x_1, \ldots, x_k$ (c'est-à-dire qu'elle est conservée par toute permutation de ces derniers).

Si la fonction est donnée aux points $x_1, \ldots, x_n$, alors

$$\begin{array}{llll} f(x_1) & & & \\ f(x_2) & f(x_1; x_2) & & \\ f(x_3) & f(x_2; x_3) & f(x_1; x_2; x_3) & \\ \vdots & \vdots & \vdots & \ddots \quad f(x_1; \ldots; x_n) \\ f(x_n) & f(x_{n-1}; x_n) & & \end{array}$$

est le *tableau de ses différences divisées.*

§ 5. Formule d'interpolation de Newton par différences divisées

Moyennant les différences divisées on obtient une autre expression du polynôme d'interpolation (2.2).

On a l'égalité

$$f(x)-L_n(x)=f(x)-\sum_{i=1}^{n} f(x_i)\prod_{j\neq i}\frac{x-x_j}{x_i-x_j}=$$

$$=\prod_{i=1}^{n}(x-x_i)\left(\frac{f(x)}{\prod\limits_{i=1}^{n}(x-x_i)}+\sum_{i=1}^{n}\frac{f(x_i)}{(x_i-x)\prod\limits_{j\neq i}(x_i-x_j)}\right).$$

Comparant avec (4.3) on s'assure que la parenthèse n'est autre que $f(x; x_1; \ldots; x_n)$. On peut donc écrire

$$f(x) - L_n(x) = f(x; x_1; \ldots; x_n)\,\omega_n(x). \tag{1}$$

Soit $L_m(x)$ le polynôme d'interpolation de Lagrange basé sur les points $x_1, \ldots, x_m$. $L_n(x)$ se met alors sous la forme

$$L_n(x) = L_1(x) + (L_2(x) - L_1(x)) + \ldots + (L_n(x) - L_{n-1}(x)). \tag{2}$$

La différence $L_m(x) - L_{m-1}(x)$ est un polynôme de degré $m-1$ s'annulant aux points $x_1, \ldots, x_{m-1}$ puisque $L_{m-1}(x_j) = L_m(x_j) = f(x_j)$ lorsque $1 \leqslant j \leqslant m-1$; par conséquent,

$$L_m(x) - L_{m-1}(x) = A_{m-1}\omega_{m-1}(x),\ \omega_{m-1}(x) = (x - x_1)\ldots(x - x_{m-1}).$$

Posant $x = x_m$, on obtient

$$f(x_m) - L_{m-1}(x_m) = A_{m-1}\omega_{m-1}(x_m).$$

D'autre part, faisant $n = m-1$ et $x = x_m$ dans (1) il vient

$$f(x_m) - L_{m-1}(x_m) = f(x_m; x_1; \ldots; x_{m-1})\,\omega_{m-1}(x_m).$$

Ainsi, $A_{m-1} = f(x_1; \ldots; x_m)$, et

$$L_m(x) - L_{m-1}(x) = f(x_1; \ldots; x_m)\,\omega_{m-1}(x);$$

portant ces quantités dans (2), on trouve

$$L_n(x) = f(x_1) + f(x_1; x_2)(x - x_1) + \ldots$$
$$\ldots + f(x_1; \ldots; x_n)(x - x_1)\ldots(x - x_{n-1}). \tag{3}$$

C'est l'expression du *polynôme d'interpolation de Newton au moyen des différences divisées*. En comparant (1) et (3.1) il vient

$$f(x; x_1; \ldots; x_n) = \frac{f^{(n)}(\zeta)}{n!}, \quad \text{où} \quad y_1 \leqslant \zeta \leqslant y_2. \tag{4}$$

En particulier, si $f(x)$ est un polynôme

$$P_l x = \sum_{j=0}^{l} b_j x^j$$

de degré $l \leqslant n$, cette formule donne

$$P_l(x_0; \ldots; x_n) = \begin{cases} b_n & \text{pour } l = n, \\ 0 & \text{pour } l < n \end{cases}$$

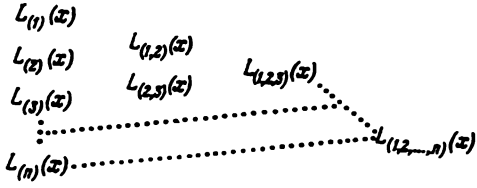
quels que soient $x_0, \ldots, x_n$.

Pour simplifier le calcul du polynôme d'interpolation, on utilisera le *procédé d'Aitken*.

Soit $L_{(k,k+1,\ldots,l)}(x)$ un polynôme d'interpolation qui prend ses valeurs aux $x_k, \ldots, x_l$, en particulier $L_{(k)}(x) = f(x_k)$. On a les égalités suivantes:

$$L_{(k,\,k+1,\,\ldots,\,l+1)}(x) = \frac{L_{(k+1,\,\ldots,\,l+1)}(x)(x-x_k) - L_{(k,\,\ldots,\,l)}(x)(x-x_{l+1})}{x_{l+1}-x_k}, \tag{5}$$

en effet, le second membre est un polynôme de degré $l-k+1$ qui se confond avec $f(x)$ aux points $x_k, \ldots, x_{l+1}$. Dans le procédé d'Aitken, on trouve $L_{(1,\ldots,n)}(x)$ en calculant successivement à l'aide de la formule (5), les éléments du tableau de valeurs des polynômes d'interpolation



On se sert volontiers de ce schéma dans le problème suivant:

Soit le tableau de valeurs d'une fonction $f(x)$; on demande de déterminer au mieux, pour chaque x, la valeur de $f(x)$ en interpolant les valeurs données par le tableau.

Supposons x fixe. Numérotons les points dans l'ordre de croissance de $|x_i - x|$. Désignons comme d'habitude $L_{(1,\ldots,m)}(x)$ par $L_m(x)$.

Plus haut nous avons obtenu la représentation (1) de l'erreur

$$f(x) - L_m(x) = f(x; x_1; \ldots; x_m)\, \omega_m(x),$$

et l'égalité

$$L_{m+1}(x) - L_m(x) = f(x_1; \ldots; x_{m+1})\,\omega_m(x). \tag{6}$$

Vu que pour $|x - x_k|$ petits,

$$f(x; x_1; \ldots, x_m) \approx \frac{f^m(x)}{m!} \approx f(x_1; \ldots; x_{m+1}),$$

il en résulte

$$L_{m+1}(x) - L_m(x) \approx f(x) - L_m(x). \tag{7}$$

Aussi la quantité $\varepsilon_m = |L_{m+1}(x) - L_m(x)|$ est-elle une estimation approchée de l'erreur dans la formule d'interpolation $f(x) \approx \approx L_m(x)$. On calcule à tour de rôle $L_0(x)$, $L_1(x)$, ε_1, $L_2(x)$, ε_2, ...; si, pour un certain m, on a $\varepsilon_m \leqslant E$, E étant la précision voulue, on arrête les calculs et on pose $f(x) \approx L_m(x)$. Si cette inégalité n'est vérifiée pour aucun m, on cherche $\varepsilon_{m_0} = \min\limits_m \varepsilon_m$ et on fait $f(x) \approx L_{m_0}(x)$; dans le cas où ce minimum est atteint pour plusieurs m, on en choisit le plus petit. Il arrive souvent qu'à partir d'un certain m, les quantités ε_m ont tendance à augmenter, ce qui fait abandonner sur le coup le calcul de $L_m(x)$, ε_m.

Ce choix de m de même que la façon de prendre les points x_i dans l'ordre des distances à x ne sont pas optimaux.

Considérons la fonction $f(x) = \dfrac{1}{x-h}$, avec h complexe.

PROBLÈME. Démontrer que pour cette fonction on a

$$f(x) - L_n(x) = \frac{1}{x-h}\left(\frac{x-x_1}{h-x_1}\right)\ldots\left(\frac{x-x_n}{h-x_n}\right).$$

Il en découle qu'en approchant la fonction par un polynôme de degré n, le maximum de précision est réalisé lorsque les points x_i sont choisis dans l'ordre de croissance de $|(x - x_i)/(h - x_i)|$. La quantité $|(x - x_i)/(h - x_i)|$ est inférieure à 1 pour les points à gauche de la droite AB: $|x - x_i| = |h - x_i|$ (fig. 2.5.1), et supérieure à cette valeur pour les points à droite de AB. Pour rendre l'erreur minimale, on prendra donc tous les points à gauche de AB et on négligera ceux à droite. Avec une disposition arbitraire des points, ces conclusions ne satisfont pas au principe de numérotage adopté ci-dessus.

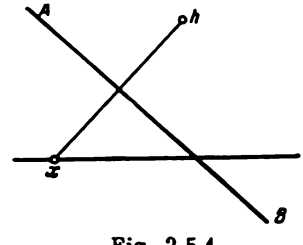


Fig. 2.5.1

Par ailleurs, il est peu probable qu'on puisse indiquer procédé d'interpolation qui soit optimal pour n'importe quelle fonction.

§ 6. Différences divisées et interpolation basée sur des points multiples

On demande de construire un polynôme $g_s(x)$ de degré $s-1$ vérifiant les conditions données:

$$\begin{aligned} &g_s(x_1)=f(x_1),\ \ldots,\ g_s^{(m_1-1)}(x_1)=f^{(m_1-1)}(x_1),\\ &\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\\ &g_s(x_n)=f(x_n),\ \ldots,\ g_s^{(m_n-1)}(x_n)=f^{(m_n-1)}(x_n); \end{aligned} \tag{1}$$

ici tous les x_i sont distincts, $s=m_1+\ldots+m_n$. On dit que $g_s(x)$ est un polynôme d'interpolation à points multiples $x_1, \ldots, x_n$ respectivement de multiplicité $m_1, \ldots, m_n$.

$g_s(x)$ se définit de façon unique. En effet, menons la démonstration par l'absurde et supposons qu'il existe deux polynômes de degré $s-1$ vérifiant les conditions (1). Alors leur différence $Q_s(x)$ est telle que

$$\begin{aligned} Q_s(x_1)=\ldots=Q_s^{(m_1-1)}(x_1)&=0,\ \ldots,\\ Q_s(x_n)=\ldots=Q_s^{(m_n-1)}(x_n)&=0; \end{aligned}$$

et les points $x_1, \ldots, x_n$ sont les zéros de $Q_s(x)$, de multiplicité respectivement $m_1, \ldots, m_n$. Le polynôme $Q_s(x) \not\equiv 0$ de degré $s-1$ possède donc s zéros, par conséquent, $Q_s(x) \equiv 0$.

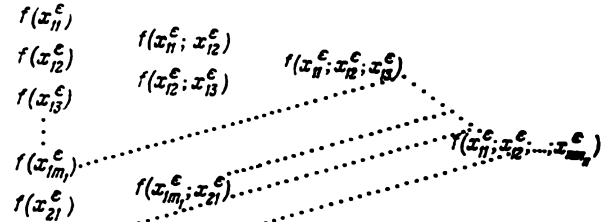
Supposons ensuite que $f(x)$ soit s fois continûment dérivable.

On démontre l'existence du polynôme d'interpolation $g_s(x)$ vérifiant les conditions (1) en l'obtenant sous une forme explicite.

On se fixe une suite d'ensembles de points x_{ij}^{ε}, $0<\varepsilon<\varepsilon_0$, $i=1, \ldots, n$, $j=1, \ldots, m_i$, qui soient tous distincts pour $0<\varepsilon<\varepsilon_0$. $x_{ij}^{\varepsilon} \to x_i$ lorsque $\varepsilon \to 0$. On pose en particulier

$$x_{ij}^{\varepsilon}=x_i+(j-1)\,\varepsilon.$$

Formons le polynôme d'interpolation $g_s^{\varepsilon}(x)$ de degré $s-1$ qui coïncide avec $f(x)$ en x_{ij}^{ε}. Le tableau des différences divisées associé à cet ensemble de points a la forme



(2)

Ecrivons la formule d'interpolation de Newton par différences divisées

$$g_s^\varepsilon(x) = A_0^\varepsilon + A_1^\varepsilon (x - x_{11}^\varepsilon) + A_2^\varepsilon (x - x_{11}^\varepsilon)(x - x_{12}^\varepsilon) + \ldots + \\ + A_{s-1}^\varepsilon (x - x_{11}^\varepsilon) \ldots (x - x_{nm_n-1}^\varepsilon),$$

où

$$A_0^\varepsilon = f(x_{11}^\varepsilon), \quad A_1^\varepsilon = f(x_{11}^\varepsilon ; x_{12}^\varepsilon),$$

$$A_2^\varepsilon = f(x_{11}^\varepsilon ; x_{12}^\varepsilon ; x_{13}^\varepsilon), \ \ldots, \quad A_{s-1}^\varepsilon = f(x_{11}^\varepsilon ; x_{12}^\varepsilon ; \ldots ; x_{nm_n}^\varepsilon).$$

En exprimant les différences en fonction des dérivées il vient

$$f(x_{il}^\varepsilon ; \ldots ; x_{im}^\varepsilon) = \frac{f^{(m-l)}(\bar{x}_{ilm}^\varepsilon)}{(m-l)!}.$$

Le passage à la limite pour $\varepsilon \to 0$ donne

$$\lim_{\varepsilon \to 0} f(x_{il}^\varepsilon ; \ldots ; x_{im}^\varepsilon) = \frac{f^{(m-l)}(x_l)}{(m-l)!}. \tag{3}$$

Ainsi, toutes les différences de la forme $f(x_{il}^\varepsilon ; \ldots ; x_{im}^\varepsilon)$ du tableau (2) tendent, pour $\varepsilon \to 0$, vers des limites qui se notent naturellement $f(\underbrace{x_l ; \ldots ; x_l}_{m-l+1 \text{ fois}})$; (il résulte de (3) que la valeur de la limite ne dépend que de $m - l$ et x_l). On a

$$f(\underbrace{x_l ; \ldots ; x_l}_{p+1 \text{ fois}}) = \frac{f^{(p)}(x_l)}{p!}. \tag{4}$$

On démontre par récurrence sur l'ordre de la différence que toutes les différences de notre tableau possèdent des limites. La propriété est immédiate pour les différences d'ordre zéro coïncidant avec les valeurs de la fonction. Supposons que cette propriété a été démontrée pour les différences d'ordre $q - 1$.

Exprimons chacune des différences d'ordre q moyennant les différences d'ordre $q - 1$:

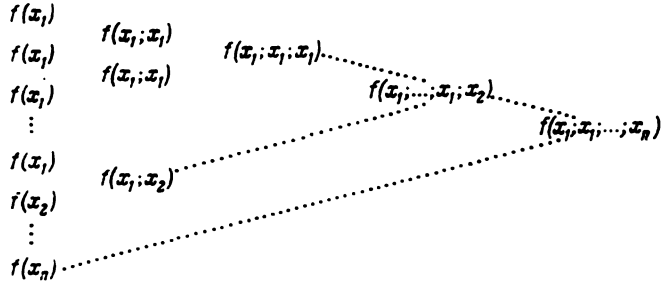
$$f(x_{il}^\varepsilon ; \ldots ; x_{jp}^\varepsilon) = \frac{f(x_{i,\,l+1}^\varepsilon ; \ldots ; x_{jp}^\varepsilon) - f(x_{il}^\varepsilon ; \ldots ; x_{j,\,p-1}^\varepsilon)}{x_{jp}^\varepsilon - x_{il}^\varepsilon}. \tag{5}$$

Pour fixer les idées, nous nous plaçons dans le cas $l < m_i$, $1 < p$. Quand $j = i$, l'existence de la limite des différences d'ordre q découle de (3). Quand $j \neq i$, la limite du dénominateur dans le second membre de (5) est $x_j - x_i \neq 0$ et celle du numérateur existe par hypothèse. Le premier membre admet donc aussi une limite. Ainsi, nous avons démontré que toutes les différences d'ordre q tendent

vers des limites, donc il en est de même pour toutes les différences du tableau. La limite du premier membre de (5) se note

$$f(\underbrace{x_i;\ \ldots;\ x_i}_{m_i-i+1 \text{ fois}};\ \underbrace{x_{i+1};\ \ldots;\ x_{i+1}}_{m_{i+1} \text{ fois}};\ \ldots;\ \underbrace{x_j;\ \ldots;\ x_j}_{p \text{ fois}}).$$

Le tableau de limites des différences



peut être rempli moyennant la formule (4) et la relation de récurrence, conséquence de (5):

$$f(\underbrace{x_i;\ \ldots;\ x_i}_{m_i-i+1 \text{ fois}};\ \ldots;\ \underbrace{x_j;\ \ldots;\ x_j}_{p \text{ fois}}) =$$

$$= \frac{f(\underbrace{x_i;\ \ldots;\ x_i}_{m_i-i \text{ fois}};\ \ldots;\ \underbrace{x_j;\ \ldots;\ x_j}_{p \text{ fois}}) - f(\underbrace{x_i;\ \ldots;\ x_i}_{m_i-i+1 \text{ fois}};\ \ldots;\ \underbrace{x_j;\ \ldots;\ x_j}_{p-1 \text{ fois}})}{x_j - x_i} \qquad (6)$$

pour $j \neq i$.

Il en résulte que les coefficients A_k^ε tendent, pour $\varepsilon \to 0$, vers les limites A_k avec

$$A_0 = f(x_1), \quad A_1 = f'(x_1),$$

$$A_2 = \frac{f''(x_1)}{2}, \ \ldots, \ A_{s-1} = f(x_1;\ \ldots;\ x_1;\ \ldots;\ x_n;\ \ldots;\ x_n).$$

Ainsi, lorsque $\varepsilon \to 0$, les polynômes $g_s^\varepsilon(x)$ tendent vers un certain polynôme

$$g_s(x) = A_0 + A_1(x-x_1) + A_2(x-x_1)^2 + \ldots$$

$$\ldots + A_{s-1}(x-x_1)^{m_1} \ldots (x-x_{n-1})^{m_{n-1}}(x-x_n)^{m_n-1} =$$

$$= f(x_1) + f(x_1;\ x_1)(x-x_1) + f(x_1;\ x_1;\ x_1)(x-x_1)^2 + \ldots \qquad (7)$$

$g_s(x)$ s'écrit

$$g_s(x) = \sum_{i=1}^{m_1} \frac{f^{(i-1)}(x_1)}{(i-1)!}(x-x_1)^{i-1} + O((x-x_1)^{m_1}).$$

On voit qu'il vérifie les conditions imposées au point x_1. Le polynôme d'interpolation étant unique, $g_s^\varepsilon(x)$ ne varie pas par la permutation $x_1 = x_j$, $x_j = x_1$. Aussi le polynôme limite vérifie-t-il les conditions données en tout point x_j, i.e. c'est bien le polynôme cherché.

Le reste s'obtiendra d'une façon plus simple si les suites de x_{ij}^ε satisfont, pour $0 < \varepsilon < \varepsilon_0$, à la relation

$$x_{ij}^\varepsilon \in [y_1,\ y_2],$$

où, de même qu'au § 3,

$$y_1 = \min(x,\ x_1,\ \ldots,\ x_n),\quad y_2 = \max(x,\ x_1,\ \ldots,\ x_n).$$

Conformément à (3.1), on a

$$f(x) - g_s^\varepsilon(x) = \left(\frac{f^{(s)}(\zeta_\varepsilon)}{s!}\right)\omega_s^\varepsilon(x); \tag{8}$$

ici

$$\omega_s^\varepsilon(x) = \prod_{i=1}^{n}\prod_{j=1}^{m_i}(x - x_{ij}^\varepsilon).$$

Si $x \neq x_1, \ldots, x_n$, passons dans (8) à la limite pour ε tendant vers zéro. Puisque le premier membre possède la limite $f(x) - g_s(x)$ et

$$\lim_{\varepsilon\to 0}\omega_s^\varepsilon(x) = \omega_s(x) = \prod_{i=1}^{n}(x - x_i)^{m_i},$$

il existe $\lim\limits_{\varepsilon\to 0} f^{(s)}(\zeta_\varepsilon) = z(x)$; donc

$$f(x) - g_s(x) = \frac{z(x)}{s!}\omega_s(x). \tag{9}$$

En passant à la limite pour $\varepsilon \to 0$ dans les relations

$$M_1 = \min_{[y_1,\ y_2]} f^{(s)}(x) \leqslant f^{(s)}(\zeta_\varepsilon) \leqslant \max_{[y_1,\ y_2]} f^{(s)}(x) = M_2,$$

on obtient $M_1 \leqslant z(x) \leqslant M_2$. D'après le théorème de Rolle, il existe un point $\zeta \in [y_1,\ y_2]$ tel que $z(x) = f^{(s)}(\zeta)$.

L'égalité (9) devient

$$f(x) - g_s(x) = \frac{f^{(s)}(\zeta)}{s!}\omega_s(x). \tag{10}$$

Conformément à (5.1), on a

$$f(x) - g_s^\varepsilon(x) = f(x;\ x_{11}^\varepsilon;\ \ldots;\ x_{nm_n}^\varepsilon)\,\omega_s^\varepsilon(x).$$

Par passage à la limite ($\varepsilon \to 0$), il vient

$$f(x) - g_s(x) = f(x; x_1; \ldots; x_n)\, \omega_s(x). \tag{11}$$

La comparaison avec (10) donne

$$f(x; x_1; \ldots; x_n) = \frac{f^{(s)}(\zeta)}{s!},$$

qui reste vraie lorsqu'on passe à la limite pour $x \to x_j$, j quelconque. Il s'ensuit de ces relations que la formule (5.4)

$$f(x_1; \ldots; x_{N+1}) = \frac{f^{(N)}(\zeta)}{N!}$$

(dans des notations différentes) reste également juste pour $x_1, \ldots, x_{N+1}$ non tous distincts.

Nous venons de démontrer l'existence du polynôme d'interpolation vérifiant les conditions (1). Le problème d'interpolation pourrait également s'énoncer comme suit:

Soit un tableau de nombres a_{ij}, $i = 1, \ldots, n$, $j = 1, \ldots, m_i$. On demande de former un polynôme $g_s(x)$ de degré $s - 1$ satisfaisant aux conditions

$$g_s^{(j-1)}(x_i) = a_{ij}, \quad i = 1, \ldots, n, \quad j = 1, \ldots, m_i.$$

On a là un problème équivalent au problème initial car on peut toujours trouver une fonction $f(x)$ régulière telle que

$$f^{(j-1)}(x_i) = a_{ij}.$$

§ 7. Equations aux différences finies

On appelle *équations aux différences finies* des équations reliant les valeurs de fonctions d'arguments entiers. En analyse discrète, leur rôle est analogue à celui des équations différentielles et intégrales en analyse mathématique classique.

Examinons un cas élémentaire, à savoir une équation linéaire par rapport à une fonction inconnue de l'argument entier:

$$ly = \sum_{i=0}^{k} a_i(n)\, y(n+i) = f(n). \tag{1}$$

Cet analogue aux différences de l'équation différentielle linéaire d'ordre k

$$\bar{l}y = \sum_{i=0}^{k} b_i(x)\, y^{(i)}(x) = f(x) \tag{2}$$

porte le nom d'*équation aux différences linéaire d'ordre k.*

L'étude de cette équation, ainsi que les raisonnements du paragraphe précédent, montre que des cas exceptionnels tels que racines multiples, points multiples d'interpolation se traitent souvent par passage à la limite dans un cas ordinaire.

Les équations (1) et (2) seront examinées simultanément.

Chacune d'elles est de la forme $Ly = h$, avec L un opérateur linéaire. L'équation $Ly = 0$ est dite *homogène*; les formules

$$y = y(C_1, \ldots, C_l, n) \quad \text{et} \quad y = y(C_1, \ldots, C_l, x)$$

représentent la *solution générale* correspondante si, en substituant les valeurs des paramètres C_l, on obtient n'importe quelle solution de l'équation considérée. Si v est solution particulière de l'équation non homogène $Lv = h$, la différence $y - v$ est solution de l'équation homogène: $L(y - v) = h - h = 0$. Ainsi, la solution générale d'une équation complète est la somme d'une de ses solutions particulières et de la solution générale de l'équation homogène associée. Les solutions $y_1, \ldots, y_m$ de l'équation homogène $Ly = 0$ sont dites *linéairement dépendantes* dans le domaine de variation considéré de la variable indépendante s'il existe des $C_1, \ldots, C_m$ non tous nuls tels que $C_1y_1 + \ldots + C_my_m \equiv 0$. Dans le cas contraire, ces solutions sont *linéairement indépendantes*. Si les fonctions u_l sont solutions de $Lu = 0$, il en est de même de toute fonction $\sum_i C_iu_i$ vu que

$$L\left(\sum_i C_iu_i\right) = \sum_i C_i(Lu_i) = 0.$$

Soient, pour fixer les idées, $x \geqslant 0$ le domaine de définition de (2) et $n \geqslant 0$ celui de (1).

THÉORÈME

Si $b_k(x) \neq 0$ *pour* $x \geqslant 0$ *et* $b_j(x)$ *sont tous continus pour* $x \geqslant 0$,

Si $a_k(n) \neq 0$ *pour* $n \geqslant 0$ *et* $a_l(n)$ *sont tous bornés pour chaque* $n \geqslant 0$,

alors la solution générale de l'équation homogène $Ly = 0$ *s'écrit*

$$y = \sum_{i=1}^{k} C_iy_i,$$

où $y_1 \ldots, y_k$ *sont des solutions linéairement indépendantes de* $Ly = 0$.

DÉMONSTRATION

Selon le théorème d'existence $\overline{ly}=0$ admet une solution pour n'importe quelles conditions initiales $y(0), \ldots, y^{(k-1)}(0)$.

Conformément au théorème d'unicité, cette solution est unique.

Notons $y_i(x)$ les solutions de $\overline{ly}=0$ avec les conditions initiales

$$y_i^{(j-1)}=\delta_{i-j}, \quad i, j=1, \ldots, k.$$

Ces solutions forment un système linéairement indépendant. En effet, si

$$\sum_{i=1}^{k} C_i y_i(x) \equiv 0,$$

alors on a pour $j=1, \ldots, k$

$$0=\sum_{i=1}^{k} C_i y_i^{(j-1)}(0)= \sum_{i=1}^{k} C_i \delta_{i-j}=C_j.$$

L'équation homogène $ly=0$ peut s'écrire :

$$y(n+k)= -\sum_{i=0}^{k-1} \frac{a_i(n)}{a_k(n)} y(n+i). \qquad (3)$$

Si l'on se donne $y(0), \ldots, y(k-1)$, on calcule successivement $y(k)$, $y(k+1)$, ... à partir de (3). Ainsi, quels que soient $y(0), \ldots, y(k-1)$, l'équation $l(y)=0$ admet une solution.

Cette solution est unique puisque les valeurs d'une solution quelconque vérifient (3), or, les valeurs $y(k)$, $y(k+1)$, ... se définissent de façon unique à partir de cette équation.

Notons $y_i(n)$ les solutions de $ly=0$ avec les conditions initiales

$$y_i(j-1)=\delta_{i-j}, \quad i, j=1, \ldots, k.$$

$$\sum_{i=1}^{k} C_i y_i(n) \equiv 0,$$

alors on a pour $j=1, \ldots, k$

$$0=\sum_{i=1}^{k} C_i y_i(j-1)= \sum_{i=1}^{k} C_i \delta_{i-j}=C_j.$$

Par conséquent, dans le cas

$$\sum_{i=1}^{k} C_i y_i \equiv 0$$

tous les C_i sont forcément nuls et les fonctions $y_1, \ldots, y_k$ linéairement indépendantes.

Soit $y(x)$ une solution de $\bar{l}y = 0$. La fonction

$$z(x) = \sum_{j=1}^{k} y^{(j-1)}(0)\, y_j(x)$$

est solution de cette équation sous les conditions initiales

$$y(0), \ \ldots, \ y^{(k-1)}(0).$$

Vu l'unicité de la solution de

$\bar{l}y = 0$, on a

$$y(x) = \sum_{j=1}^{k} y^{(j-1)}(0)\, y_j(x).$$

C.Q.F.D.

Soit $y(n)$ une solution de $ly = 0$. La fonction

$$z(n) = \sum_{j=1}^{k} y(j-1)\, y_j(n)$$

est solution de cette équation sous les conditions initiales

$$y(0), \ \ldots, \ y(k-1).$$

Vu l'unicité de la solution de

$ly = 0$, on a

$$y(n) = \sum_{j=1}^{k} y(j-1)\, y_j(n).$$

C.Q.F.D.

Dans les cours d'équations différentielles, on démontre le théorème suivant. Si l'on connaît k solutions linéairement indépendantes de l'équation homogène $\bar{l}y = 0$, on trouve la solution de l'équation non homogène (2) en résolvant

$$dC_j/dx = g_j(x),$$

avec $g_j(x)$ connues. De même, connaissant k solutions linéairement indépendantes de l'équation homogène $ly = 0$, on ramène la résolution de l'équation complète à celle des équations aux différences simples

$$C_j(n+1) - C_j(n) = g_j(n),$$

avec $g_j(n)$ connues. L'intérêt pratique de ces théorèmes étant minime, nous nous permettons de passer outre.

On se propose maintenant d'étudier les équations à coefficients constants

$$\bar{l}y = \sum_{i=0}^{k} b_i y^{(i)}(x) = f(x), \quad b_k \neq 0 \qquad (4)$$

$$ly = \sum_{i=0}^{k} a_i y(n+i) = f(n), \quad a_k \neq 0 \qquad (5)$$

et les équations homogènes correspondantes

$$\bar{l}y = \sum_{i=0}^{k} b_i y^{(i)}(x) = 0. \qquad (6)$$

$$ly = \sum_{i=0}^{k} a_i y(n+i) = 0. \qquad (7)$$

Recherchons les solutions particulières de l'équation homogène.

Portant dans (6) la solution particulière supposée égale à $\exp(\lambda x)$, on obtient

$$\Big(\sum_{i=0}^{k} b_i \lambda^i\Big) \exp(\lambda x) = 0.$$

Dans le cas de l'équation (7), on écrira la fonction $\exp(\lambda n)$ sous la forme μ^n, $\mu = \exp\lambda$. Portant dans (7) on obtient

$$\Big(\sum_{i=0}^{k} a_i \mu^i\Big) \mu^n = 0.$$

Ainsi, à chaque racine de l'équation

$$\sum_{i=0}^{k} b_i \lambda^i = 0. \qquad (8)$$

$$\sum_{i=0}^{k} a_i \mu^i = 0, \qquad (9)$$

appelée *équation caractéristique*, correspond la solution particulière

$$\exp(\lambda x).$$

$$\mu^n.$$

Si toutes les racines de l'équation caractéristique sont simples, on obtient k solutions distinctes. Montrons qu'il correspond à toute racine l-uple de cette équation l solutions distinctes de l'équation homogène .

$$\exp(\lambda x),\ x\exp(\lambda x),\ \ldots,\ x^{l-1}\exp(\lambda x).$$

$$\mu^n,\ C_n^1\mu^{n-l_1},\ \ldots,\ C_n^{l-1}\mu^{n-l+1}.$$

Soit, pour fixer les idées, $\lambda_1 = \ldots = \lambda_l$, $\mu_1 = \ldots = \mu_l$.

Décomposons le polynôme caractéristique en produit de facteurs

$$\sum_{i=0}^{k} b_i \lambda^i = b_k \prod_{j=1}^{k} (\lambda - \lambda_j),$$

$$\sum_{i=0}^{k} a_i \mu^i = a_k \prod_{j=1}^{k} (\mu - \mu_j),$$

et se fixons un paramètre réel $\varepsilon > 0$, $\varepsilon \to 0$.

Prenons $\lambda_{j\varepsilon}$ tels que
a) ils sont distincts pour $j = 1, \ldots, l$,
b) ils tendent vers λ_j quand $\varepsilon \to 0$, quels que soient $j \leqslant k$.

Prenons $\mu_{j\varepsilon}$ tels que
a) ils sont distincts pour $j = 1, \ldots, l$,
b) ils tendent vers μ_j quand $\varepsilon \to 0$, quels que soient $j \leqslant k$.

Formons les équations caractéristiques associées à ces racines

$$0 = b_k \prod_{j=1}^{k} (\lambda - \lambda_{j\varepsilon}) = \sum_{i=0}^{k} b_{i\varepsilon} \lambda^i.$$

$$0 = a_k \prod_{j=1}^{k} (\mu - \mu_{j\varepsilon}) = \sum_{i=0}^{k} a_{i\varepsilon} \mu^i.$$

On a de toute évidence

$$b_{i\varepsilon} \to b_i$$

$$a_{i\varepsilon} \to a_i \qquad (10)$$

pour $\varepsilon \to 0$.

A ces équations caractéristiques sont associées les équations

$$\sum_{i=0}^{k} b_{i\varepsilon} y_\varepsilon^i x = 0. \qquad (11)$$

$$\sum_{i=0}^{k} a_{i\varepsilon} y_\varepsilon (n+i) = 0. \qquad (12)$$

Admettons qu'avec un ε supérieur à zéro, nous pouvons trouver les solutions:

$y_\varepsilon(x)$ de l'équation (11) telles que pour tout $x \geqslant 0$, il existe la limite

$$\lim_{\varepsilon \to 0} y_\varepsilon(x) = Y(x),$$

et $y_\varepsilon(x)$ converge, avec toutes ses dérivées jusqu'à l'ordre k inclus, vers $Y(x)$ uniformément sur tout segment fini $[x_1, x_2]$.

En passant à la limite dans (11) et en tenant compte de (10), on obtient que la fonction limite $Y(x)$ satisfait à l'équation (6).

$y_\varepsilon(n)$ de l'équation (12) telles que pour tout $n \geqslant 0$, il existe la limite

$$\lim_{\varepsilon \to 0} y_\varepsilon(n) = Y(n).$$

En passant à la limite dans (12) et en tenant compte de (10), on obtient que la fonction limite $Y(n)$ satisfait à l'équation (7).

Formons les suites de $y_\varepsilon(x)$ et de $y_\varepsilon(n)$ qui convergent vers les solutions particulières de (6) et (7) associées aux racines multiples. On utilisera avec succès les différences divisées. Commençons par le cas d'une racine double. Posons

$$y_{2\varepsilon}(x) = (\exp \lambda x)(\lambda_{1\varepsilon};\ \lambda_{2\varepsilon}) = \frac{\exp(\lambda_{2\varepsilon}x) - \exp(\lambda_{1\varepsilon}x)}{\lambda_{2\varepsilon} - \lambda_{1\varepsilon}}.$$

$$y_{2\varepsilon}(n) = (\mu^n)(\mu_{1\varepsilon};\ \mu_{2\varepsilon}) = \frac{\mu_{2\varepsilon}^n - \mu_{1\varepsilon}^n}{\mu_{2\varepsilon} - \mu_{1\varepsilon}}.$$

Ces fonctions sont visiblement solutions des équations (11), (12). Récrivons-les comme suit:

$$y_{2\varepsilon}(x) = x \exp(\lambda_{1\varepsilon}x) \times \times \frac{\exp((\lambda_{2\varepsilon} - \lambda_{1\varepsilon})x) - 1}{(\lambda_{2\varepsilon} - \lambda_{1\varepsilon})x}.$$

$$y_{2\varepsilon}(n) = \mu_{2\varepsilon}^{n-1} + \mu_{2\varepsilon}^{n-2}\mu_{1\varepsilon} + \ldots + \mu_{1\varepsilon}^{n-1}.$$

Par passage à la limite pour $\varepsilon \to 0$ il vient

$$y_{2\varepsilon}(x) \to x \exp(\lambda_1 x)$$

$$y_{2\varepsilon}(n) \to \mu_1^{n-1} + \ldots + \mu_1^{n-1} = n\mu_1^{n-1}.$$

Nous obtenons ainsi la deuxième solution linéairement indépendante associée à la racine double. Le cas d'une racine de multiplicité supérieure se traite de façon analogue. Pour $1 \leqslant q \leqslant l$, posons

$$y_{q\varepsilon}(x) = (\exp(\lambda x))(\lambda_{1\varepsilon}; \ldots; \lambda_{q\varepsilon}).$$

$$y_{q\varepsilon}(n) = (\mu^n)(\mu_{1\varepsilon}; \ldots; \mu_{q\varepsilon}).$$

Selon le lemme (4.3) $y_{q\varepsilon}(x)$ s'écrit

$$\sum_{j=1}^{q} c_{j\varepsilon} \exp(\lambda_{j\varepsilon}x)$$

et donc est solution de (11).

$y_{q\varepsilon}(n)$ s'écrit

$$\sum_{j=1}^{q} c_{j\varepsilon}\mu_{j\varepsilon}^n$$

et donc est solution de (12).

Considérons d'abord le cas où tous les

$$\lambda_{j\varepsilon},\quad j=1,\ldots,q, \qquad\qquad \mu_{j\varepsilon},\quad j=1,\ldots,q,$$

sont réels.

Moyennant la formule (5.4)

$$f(x_1,\ldots,x_q)=\frac{f^{q-1}(\zeta)}{(q-1)!},\quad \text{où}\quad y_1\leqslant\zeta\leqslant y_2, \tag{13}$$

on obtient

$$y_{q\varepsilon}(x)=\frac{x^{q-1}}{(q-1)!}\exp(\lambda_\varepsilon x),$$

où $\lambda_\varepsilon\in[\min(\lambda_{1\varepsilon},\ldots,\lambda_{q\varepsilon}),\ \max(\lambda_{1\varepsilon},\ldots,\lambda_{q\varepsilon})]$.

$$y_{q\varepsilon}(n)=C_n^{q-1}\mu_\varepsilon^{n-q+1},$$

où $\mu_\varepsilon\in[\min(\mu_{1\varepsilon},\ldots,\mu_{q\varepsilon}),\ \max(\mu_{1\varepsilon},\ldots,\mu_{q\varepsilon})]$.

Lorsque $\varepsilon\to 0$,

$$y_{q\varepsilon}(x)\to Y_q(x)=\frac{x^{q-1}}{(q-1)!}\exp(\lambda_1 x). \tag{14}$$

$$y_{q\varepsilon}(n)\to Y_q(n)=C_n^{q-1}\mu_1^{n-q+1}. \tag{15}$$

Si $\lambda_{j\varepsilon}$ (resp. $\mu_{j\varepsilon}$) ne sont pas tous réels, les relations (14) (resp. (15)) restent en vigueur, mais non la démonstration puisque pour $x_1,\ldots,x_q$ complexes la formule (13) ne joue plus.

Bornons-nous à démontrer (14), (15) en choisissant spécialement $\lambda_{j\varepsilon}$, $\mu_{j\varepsilon}$. On suppose alors que $\mu_1\neq 0$; pour $\mu_1=0$, on prend $\mu_{j\varepsilon}=(j-1)\varepsilon$, et la démonstration ci-dessus reste entière. Pour $\varepsilon>0$, posons

$$\lambda_{j\varepsilon}=\lambda_1+(j-1)\varepsilon \tag{16}$$

avec $j=1,\ldots,l$.

$$\mu_{j\varepsilon}=\mu_1(1+(j-1)\varepsilon) \tag{17}$$

avec $j=1,\ldots,l$.

En utilisant (4.3) il vient

$$(\exp(\lambda x))(\lambda_{1\varepsilon};\ \ldots;\ \lambda_{q\varepsilon})=\sum_{j=1}^{q}\frac{\exp[(\lambda_1+(j-1)\varepsilon)x]}{\prod\limits_{i\neq j}((\lambda_1+(j-1)\varepsilon)-(\lambda_1+(i-1)\varepsilon))}=$$

$$=\left(\sum_{j=1}^{q}\frac{\exp(j-1)\varepsilon x}{\prod\limits_{i\neq j}((j-1)\varepsilon-(i-1)\varepsilon)}\right)\exp(\lambda_1 x)=$$

$$=((\exp x)(0;\ \varepsilon;\ \ldots;\ (q-1)\varepsilon))\exp(\lambda_1 x);$$

de même

$$(\mu^n)(\mu_{1\varepsilon};\ \ldots;\ \mu_{q\varepsilon})=\sum_{j=1}^{q}\frac{\mu_1^n(1+(j-1)\varepsilon)^n}{\prod\limits_{i\neq j}(\mu_1(1+(j-1)\varepsilon)-\mu_1(1+(i-1)\varepsilon))}=$$

$$=\left(\sum_{j=1}^{q}\frac{(1+(j-1)\varepsilon)^n}{\prod\limits_{i\neq j}((1+(j-1)\varepsilon)-(1+(i-1)\varepsilon))}\right)\mu_1^{n-q+1}=$$

$$=(\mu^n(1;\ 1+\varepsilon;\ \ldots;\ 1+(q-1)\varepsilon)\,\mu_1^{n-q+1}.$$

D'où, vu (13),

$$(\exp(\lambda x))(\lambda_{1\varepsilon}; \ldots; \lambda_{q\varepsilon}) = \frac{x^{q-1}}{(q-1)!}\exp((\lambda_1+\bar{\lambda}_\varepsilon)x),$$

où

$$0 \leqslant \bar{\lambda}_\varepsilon \leqslant (q-1)\varepsilon.$$

$$(\mu^n)(\mu_{1\varepsilon}; \ldots; \mu_{q\varepsilon}) = C_n^{q-1}\mu_1^{n-q+1}\bar{\mu}_\varepsilon^{n-q+1}$$

où

$$1 \leqslant \bar{\mu}_\varepsilon \leqslant 1+(q-1)\varepsilon.$$

Il en découle que pour $\varepsilon \to 0$, on a les relations

$$(\exp(\lambda x))(\lambda_{1\varepsilon}; \ldots; \lambda_{q\varepsilon}) \to Y_q(x), \quad | \quad (\mu^n)(\mu_{1\varepsilon}; \ldots; \mu_{q\varepsilon}) \to Y_q(n),$$

q pouvant être égal à 1, ..., l.

On voit que dans le cas de l'équation (7) tout comme dans (6) on a pu associer à chaque racine l-uple de l'équation caractéristique l solutions différentes.

THÉORÈME. *L'ensemble de toutes les solutions associées aux racines de l'équation caractéristique* (9) *forme un système fondamental* (i.e. elles sont linéairement indépendantes et la solution de (7) peut être obtenue sous forme de leur combinaison linéaire).

En voici une démonstration sommaire. Soient $y_1(n), \ldots, y_k(n)$ des solutions particulières de (7).

LEMME. *Le système de solutions* $y_1(n), \ldots, y_k(n)$ *est fondamental si*

$$\det \| y_i(j-1) \| \neq 0.$$

En effet, supposons que ces fonctions sont linéairement dépendantes. Il existe alors des C_i non tous nuls tels que

$$\sum_{i=1}^{k} C_i y_i(n) = 0$$

quels que soient n, en particulier pour $n = 0, \ldots, k-1$. Or, la condition $\det \| y_i(j-1) \| \neq 0$ implique que le système d'équations linéaires

$$\sum_{i=1}^{k} C_i y_i(j-1) = 0, \quad j = 1, \ldots, k,$$

n'admet que la solution triviale $C_1 = \ldots = C_k = 0$, ce qui contredit notre hypothèse. Soient $y_1(n), \ldots, y_k(n)$ des solutions vérifiant la condition du lemme. Montrons que toute solution de (7) est combinaison linéaire de ces solutions. Le système

$$\sum_{i=1}^{k} C_i y_i(j-1) = y(j-1), \quad j = 1, \ldots, k,$$

admet une solution pour n'importe quels $y(0), \ldots, y(k-1)$. Ainsi, on peut indiquer pour $y(n)$ quelconque des $C_1, \ldots, C_k$ tels que la solution de

l'équation homogène

$$z(n)=\sum_{i=1}^{k} C_i y_i(n)$$

coïncide avec $y(n)$ pour $n=0, \ldots, k-1$. L'unicité de la solution d'une équation aux différences réalisant les conditions initiales $y(0), \ldots, y(k-1)$ entraîne

$$y(n)=z(n)=\sum_{i=1}^{k} C_i y_i(n)$$

quels que soient n. Le lemme est démontré.

Si l'équation caractéristique (9) a toutes ses racines distinctes, on a, pour $y_i(n)=\mu_i^n$,

$$\det \| y_i(j-1) \| = \prod_{i>j} (\mu_i - \mu_j) \neq 0.$$

Considérons le cas général où certaines racines de (9) sont multiples: $\mu_1 = \ldots = \mu_{l_1}$, $\mu_{l_1+1} = \ldots = \mu_{l_1+l_2}, \ldots$; μ_1, μ_{l_1+1}, $\mu_{l_1+l_2+1}$, ... sont distinctes. Ecrivons des solutions particulières de l'équation aux différences perturbées (12):

$$y_{1\varepsilon}(n)=(\mu^n)(\mu_{1\varepsilon}), \quad y_{l_1\varepsilon}(n)=(\mu^n)(\mu_{1\varepsilon}; \ldots; \mu_{l_1\varepsilon}),$$

$$y_{l_1+1,\,\varepsilon}(n)=(\mu^n)(\mu_{l_1+1,\,\varepsilon}), \ldots, y_{l_1+l_2,\,\varepsilon}(n)=(\mu^n)(\mu_{l_1+1,\,\varepsilon}; \ldots; \mu_{l_1+l_2,\,\varepsilon}), \ldots$$

On a vu que le choix spécial (17) des racines de l'équation perturbée fait $y_{j\varepsilon}(n) \to Y^{(j)}(n)$ pour $\varepsilon \to 0$ avec

$$\begin{aligned} &Y^{(1)}(n)=\mu_1^n, && Y^{(2)}(n)=C_n^1 \mu_1^{n-1}, \ldots, \\ &Y^{(l_1)}(n)=C_n^{l_1-1}\mu_1^{n-l_1+1}, && Y^{(l_1+1)}(n)=\mu_{l_1+1}^n, \ldots \end{aligned} \tag{18}$$

Il est clair que $W_\varepsilon = \det \| y_{j\varepsilon}(j-1) \| \to \det \| Y^{(i)}(j-1) \| = W$. Transformant linéairement les lignes de la matrice $\| y_{i\varepsilon}(j-1) \|$ on a

$$|W_\varepsilon| = \Big| \prod_{i>j}' (\mu_{i\varepsilon} - \mu_{j\varepsilon}) \Big| ;$$

l'accent du signe de produit signifie l'absence de facteurs $(\mu_{i\varepsilon} - \mu_{j\varepsilon})$ tels que, pour $\varepsilon \to 0$, $\mu_{i\varepsilon}$ et $\mu_{j\varepsilon}$ tendent vers une même limite, Il en résulte

$$W = \lim_{\varepsilon \to 0} W_\varepsilon \neq 0,$$

et le système de fonctions $Y^{(j)}(n)$ est fondamental, c.q.f.d.

THÉORÈME. *L'ensemble des combinaisons linéaires*

$$\sum_{q=1}^{l} c_q C_n^{q-1} \mu_1^{n-q+1} \tag{19}$$

des solutions particulières de (7) *associées à la racine* μ_1 *de multiplicité* l *coïncide avec l'ensemble des fonctions*

$$P_{l-1}(n)\,\mu_1^n, \tag{20}$$

où $P_{l-1}(n)$ *sont des polynômes arbitraires de degré* $l-1$.

Puisque toute fonction $C_n^{q-1}\mu_1^{1-q}$ est un polynôme de degré $q-1<l$ en n, toute fonction de la forme (19) est représentable sous la forme (20). D'autre part, soit $P_{l-1}(n)$ un polynôme de degré $l-1$ en n. Chaque polynôme de degré $l-1$ est un polynôme d'interpolation du même degré pour lui-même quels que soient l points d'interpolation. Aussi peut-on faire $n=l$, $L_n = P_{l-1}$, $f = P_{l-1}$ dans (5.3). Posons de plus $x_j = j-1$ et $x = n$; on obtient

$$P_{l-1}(n) = A_0 + A_1 n + A_2 n(n-1) + \ldots + A_{l-1} n(n-1)\ldots \\ \ldots (n-(l-2)),$$

où $A_j = P_{l-1}(0; \ldots; j)$; cette égalité se récrit

$$P_{l-1}(n) = \sum_{q=0}^{l-1} c_q C_n^{q-1}\mu_1^{1-q}, \quad c_q = A_q(q-1)!\,\mu_1^{q-1};$$

il s'ensuit que toute solution du type (20) peut se mettre sous la forme (19). Le théorème est démontré.

Ainsi, le système fondamental (18) peut être remplacé par

$$Y_1(n) = \mu_1^n,\ Y_2(n) = n\mu_1^n, \ldots, Y_{l_1}(n) = n^{l_1-1}\mu_1^n,$$
$$Y_{l_1+1}(n) = \mu_{l_1+1}^n, \ldots,$$

forme qui s'avère souvent plus commode.

Considérons maintenant l'équation aux différences

$$\sum_{i=0}^{k} a_i y(n+i) = \Big(\sum_{j=0}^{m-1} C_j n^j\Big)\sigma^n. \tag{21}$$

Soit σ une racine de multiplicité l de l'équation caractéristique (9); en particulier, si σ n'est pas racine de cette équation, $l=0$. On démontre que (21) a une solution particulière de la forme

$$\Big(\sum_{j=l}^{l+m-1} d_j n^j\Big)\sigma^n; \tag{22}$$

d_j sont trouvés par la méthode des coefficients indéterminés.

§ 8. Polynômes de Tchebycheff

Les polynômes de Tchebycheff $T_n(x)$ $(n \geqslant 0)$ sont définis par les relations

$$T_0(x) = 1,\ T_1(x) = x,$$
$$T_{n+1}(x) = 2xT_n(x) - T_{n-1}(x) \text{ pour } n > 0. \tag{1}$$

En utilisant la formule de récurrence (1) on obtient, par exemple,

$$T_2(x) = 2x^2 - 1, \qquad T_3(x) = 4x^3 - 3x,$$
$$T_4(x) = 8x^4 - 8x^2 + 1, \quad T_5(x) = 16x^5 - 20x^3 + 5x, \ldots$$

Le terme de plus haut degré $T_{n+1}(x)$ s'obtient du terme précédent $T_n(x)$ en multipliant celui-ci par $2x$, donc, pour $n > 0$, $T_n(x) = 2^{n-1}x^n$.

Les polynômes $T_{2n}(x)$ sont des fonctions paires et les polynômes $T_{2n+1}(x)$ des fonctions impaires.

La propriété est immédiate pour $n = 0$. Supposons qu'elle est valable pour un certain n. Nous obtenons alors: $2xT_{2n+1}(x)$ (fonction paire) et, conformément à (1), $T_{2n+2}(x)$ (fonction paire). Alors $2xT_{2n+2}(x)$ est impaire et $T_{2n+3}(x)$, impaire selon (1).

On a pour tout θ

$$\cos((n+1)\theta) = 2\cos\theta\cos n\theta - \cos((n-1)\theta).$$

Posant $\theta = \arccos x$ il vient

$$\cos((n+1)\arccos x) = 2x\cos(n\arccos x) - \cos((n-1)\arccos x).$$

La fonction $\cos(n\arccos x)$ vérifie la même équation aux différences par rapport à la variable n (1) que $T_n(x)$; pour $n = 0$ et $n = 1$, les conditions initiales sont les mêmes:

$$\cos(0\cdot\arccos x) = 1 = T_0(x),$$
$$\cos(1\cdot\arccos x) = x = T_1(x);$$

aussi pour tout n

$$T_n(x) = \cos(n\arccos x). \tag{2}$$

Donc,

$$|T_n(x)| \leqslant 1 \quad \text{si} \quad |x| \leqslant 1. \tag{3}$$

$|T_n(x)|$ n'est forcément inférieur ou égal à 1 pour chaque x. Si $|x| > 1$, alors $\arccos x$ n'est pas réel, et le cosinus de ce nombre est supérieur à 1.

La relation de récurrence (1) est une équation aux différences d'équation caractéristique

$$\mu^2 - 2x\mu + 1 = 0,$$

d'où

$$\mu_{1,2} = x \pm \sqrt{x^2-1}.$$

Pour $x \neq \pm 1$, les racines sont simples, donc $T_n(x) = C_1\mu_1^n + C_2\mu_2^n$. Les conditions initiales $T_0(x) = 1$, $T_1(x) = x$ donnent $C_1 = C_2 = 1/2$; de cette façon

$$T_n(x) = \frac{(x+\sqrt{x^2-1})^n + (x-\sqrt{x^2-1})^n}{2}. \tag{4}$$

Pour $x = z$ complexes, la fonction $\sqrt{z^2 - 1}$ est multivoque. On sous-entendra par $\sqrt{z^2 - 1}$ la branche qui prend des valeurs positives pour les $z > 1$ réels et est définie dans le plan muni de la coupure $[-1, 1]$.

L'équation

$$T_n(x) = \cos(n \arccos x) = 0$$

entraîne que

$$x_m = \cos\left(\frac{\pi(2m+1)}{2n}\right), \quad m = 0, \ldots, n-1,$$

sont racines de $T_n(x)$. Vu (2), (3), les points d'extrémum de $T_n(x)$ sur $[-1, 1]$ sont les points en lesquels $|T_n(x)| = 1$. En résolvant cette équation, on a

$$x_{(m)} = \cos\left(\frac{\pi m}{n}\right), \quad m = 0, \ldots, n$$

et

$$T_n(x_{(m)}) = \cos \pi m = (-1)^m.$$

On dit des polynômes

$$\overline{T}_n(x) = 2^{1-n} T_n(x) = x^n + \ldots$$

qu'ils *s'écartent de zéro le moins possible*. En effet, on a le

LEMME. *Si $P_n(x)$ est un polynôme de degré n de coefficient de la plus haute puissance de x égal à 1, alors*

$$\max_{[-1, 1]} |P_n(x)| \geqslant \max_{[-1, 1]} |\overline{T}_n(x)| = 2^{1-n}. \tag{5}$$

DÉMONSTRATION. Raisonnons par l'absurde et supposons que le degré de $\overline{T}_n(x) - P_n(x)$ est $n - 1$; par ailleurs

$$\operatorname{sg}(\overline{T}_n(x_{(m)}) - P_n(x_{(m)})) = \operatorname{sg}((-1)^m 2^{1-n} - P_n(x_{(m)})) = (-1)^m,$$

puisque par hypothèse $|P_n(x_{(m)})| < 2^{1-n}$ pour tous les m. Ainsi, le polynôme $\overline{T}_n(x) - P_n(x)$ est alternativement négatif et positif entre les points $x_{(m)}$ et $x_{(m+1)}$. Le polynôme $\overline{T}_n(x) - P_n(x)$ de degré $n - 1$ et différent de zéro (puisqu'il l'est aux points $x_{(m)}$) possède n racines distinctes. Nous avons abouti à une contradiction.

PROBLÈME. Démontrer la proposition plus forte: si

$$P_n(x) = x^n + \ldots \neq \overline{T}_n(x),$$

alors

$$\max_{[-1, 1]} |P_n(x)| > 2^{1-n}.$$

Par changement linéaire $x' = \frac{b+a}{2} + \frac{b-a}{2}x$, le segment $[-1, 1]$ devient un segment $[a, b]$ donné. Le polynôme $\overline{T}_n(x)$ se transforme alors en $\overline{T}_n\left(\frac{2x-(b+a)}{b-a}\right)$ de coefficient de la plus haute puissance de x égal à $(2/(b-a))^n$. En vertu du lemme on affirme que le polynôme

$$\overline{T}_n^{[a, b]}(x) = (b-a)^n 2^{1-2n} T_n\left(\frac{2x-(b+a)}{b-a}\right)$$

dans lequel le coefficient correspondant est 1 s'écarte de zéro sur $[a, b]$ le moins possible. Cela signifie qu'on a, pour tout polynôme $P_n(x)$ de degré n dont le coefficient du terme de plus haut degré est 1, l'inégalité

$$\max_{[a, b]} |P_n(x)| \geqslant \max_{[a, b]} |\overline{T}_n^{[a, b]}(x)| = (b-a)^n 2^{1-2n}. \qquad (6)$$

On vérifie aisément que les zéros de $\overline{T}_n^{[a,b]}(x)$ sont les points

$$x_m = \frac{b+a}{2} + \frac{b-a}{2}\cos\left(\frac{\pi(2m+1)}{2n}\right), \quad m = 0, \ldots, n-1.$$

On remarquera également que pour $n \geqslant 1$, les polynômes $\widetilde{T}_0(x) = 1/\sqrt{2}$ et $\widetilde{T}_n(x) = T_n(x)$ forment sur $[-1, 1]$ un système orthonormé de poids $\frac{2}{\pi\sqrt{1-x^2}}$. Vérifions la propriété d'orthonormalité de ces fonctions pour m, $n \neq 0$. La substitution $x = \cos\theta$ donne

$$\int_{-1}^{1} \frac{2T_n(x)\,T_m(x)}{\pi\sqrt{1-x^2}}\,dx = \frac{2}{\pi}\int_0^{\pi} \cos n\theta \cos m\theta\,d\theta =$$

$$= \frac{1}{\pi}\int_0^{\pi} (\cos((n-m)\theta) + \cos((n+m)\theta))\,d\theta = \delta_{n-m} + \delta_{n+m}.$$

Le deuxième terme s'annule pour $n, m \geqslant 0$ dès que $n^2 + m^2 \neq 0$. D'où l'égalité cherchée

$$\int_{-1}^{1} \frac{2\widetilde{T}_n(x)\,\widetilde{T}_m(x)}{\pi\sqrt{1-x^2}}\,dx = \delta_{n-m}.$$

§ 9. Minimisation de l'évaluation du reste d'une formule d'interpolation

Supposons qu'une fonction $f(x)$ soit approchée sur $[a, b]$ par un polynôme d'interpolation de degré $n-1$ basé sur les points $x_1, \ldots, x_n \in [a, b]$; convenons d'évaluer l'erreur pour la norme $\| f \| = \sup_{[a,b]} | f(x) |$. D'après (3.1), on a

$$f(x) - L_n(x) = \frac{f^{(n)}(\zeta)\, \omega_n(x)}{n!},$$

avec $\zeta \in [a, b]$ si $x \in [a, b]$. D'où l'évaluation de l'erreur d'interpolation

$$\| f - L_n \| \leqslant \frac{\| f^{(n)} \| \, \| \omega_n \|}{n!}. \tag{1}$$

Minimisons le second membre en choisissant habilement les points $x_1, \ldots, x_n$ (c'est d'ailleurs pour cette raison que nous avons introduit les polynômes de Tchebycheff s'écartant de zéro le moins possible). Le coefficient du terme de plus haut degré du polynôme $\omega_n(x) = (x - x_1) \ldots (x - x_n)$ vaut 1, donc $\| \omega_n \| \geqslant (b-a)^n 2^{1-2n}$ selon (8.6). Si l'on prend pour points d'interpolation

$$x_m = \frac{b+a}{2} + \frac{b-a}{2} \cos\left(\frac{\pi(2m+1)}{2}\right), \quad m = 0, \ldots, n-1, \tag{2}$$

alors

$$\omega_n(x) = (b-a)^n 2^{1-2n} T_n\left(\frac{2x-(b+a)}{b-a}\right)$$

et

$$\| \omega_n \| = (b-a)^n 2^{1-2n}.$$

Par conséquent, cette disposition des points permet la meilleure évaluation qui puisse être obtenue comme une conséquence de (1)

$$\| f - L_n \| \leqslant \frac{\| f^{(n)} \| \, (b-a)^n 2^{1-2n}}{n!}. \tag{3}$$

Dans (1) le maximum du produit est remplacé par le produit des maxima des facteurs, et nous pourrions escompter une estimation meilleure que (3). Or, il en va autrement ; en effet si $f(x) = a_n x^n + \ldots$ est un polynôme de degré n, alors

$$f^{(n)}(\zeta) = a_n n! = \text{const},$$

et l'inégalité (1) devient égalité. Quels que soient les points d'interpolation, on a alors

$$\| f - L_n \| \geqslant \frac{\| f^{(n)} \| \, (b-a)^n 2^{1-2n}}{n!} = a_n (b-a)^n 2^{1-2n}.$$

L'optimisation des méthodes de résolution de problèmes d'une classe déterminée constitue, nous l'avons dit, un objet majeur du calcul mathématique. Voici son idée générale.

On se donne une classe P de problèmes p à résoudre et un ensemble M de méthodes de résolution. Soit $e(p, m)$ l'erreur de méthode m sur le problème p. La quantité

$$e(P, m) = \sup_{p \in P} e(p, m)$$

est dite *erreur de méthode sur la classe P de problèmes* et la quantité

$$E(P, M) = \inf_{m \in M} e(P, m)$$

estimation optimale de l'erreur de méthode de M sur la classe P. S'il existe une méthode $m \in M$ qui réalise cette estimation, i.e. $E(P, M) = e(P, m)$, m s'appelle *méthode optimale.*

La solution du problème d'optimisation de la répartition des points d'interpolation que nous venons d'obtenir peut s'énoncer en termes analogues.

Soient P une famille de problèmes d'approximation de fonctions définies sur $[a, b]$ dans la condition $|f^{(n)}(x)| \leqslant A_n$ et M un ensemble de méthodes consistant à remplacer la fonction par son polynôme d'interpolation $L_n(x)$ basé sur les points $x_1, \ldots, x_n$, de sorte que la méthode m se définit par la donnée de $x_1, \ldots, x_n$. Soit enfin $e(p, m) = \| f - L_n \|$. On a, en vertu de (1),

$$e(P, m) \leqslant \frac{A_n \| \omega_n \|}{n!}.$$

D'autre part, l'erreur sur l'approximation du polynôme

$$P_{n+1}(x) = \frac{A_n}{n!} x^n + \ldots,$$

qui est un problème de la classe considérée, est

$$e(p, m) = \frac{A_n \| \omega_n \|}{n!};$$

par conséquent,

$$e(P, m) = \frac{A_n \| \omega_n \|}{n!}.$$

On a vu que

$$E(P, M) = \inf_m e(P, m) = \inf_{x_1, \ldots, x_n} \left(\frac{A_n \| \omega_n \|}{n!} \right) = \frac{A_n (b-a)^n 2^{1-2n}}{n!}.$$

Ainsi, l'interpolation aux points de base (2) des polynômes de Tchebycheff est un procédé optimal dans le sens considéré.

Le développement du calcul mathématique doit beaucoup à une formulation et une résolution rigoureuses de problèmes variationnels de ce type. En fait, on pourrait exposer le gros du présent cours comme résolution de divers problèmes d'optimisation. Nous nous en garderons bien afin de ne pas exagérer le rôle du problème d'optimisation des méthodes.

Pour conclure, comparons (3) avec l'évaluation de l'erreur commise en développant une fonction en série de Taylor. En vertu de nos raisonnements du § 6, la tranche

$$t_n(x) = \sum_{j=0}^{n-1} \frac{f^{(j)}\left(\frac{a+b}{2}\right)}{j!}\left(x - \frac{a+b}{2}\right)^j$$

de la série de Taylor coïncide avec le polynôme d'interpolation de Lagrange basé sur l'unique point n-uple $\frac{a+b}{2}$. Aussi l'évaluation doit-elle naturellement s'améliorer pour la répartition optimale des points (2). En effet, l'estimation

$$\| f - t_n \| \leqslant \frac{\| f^{(n)} \| (b-a)^n 2^{-n}}{n!}$$

de l'erreur sur la tranche de la série taylorienne est 2^{n-1} fois moins bonne que (3). Pour tout autre choix du point en lequel la fonction admet un développement en série de Taylor, la première estimation est plus mauvaise.

§ 10. Différences finies

Supposons les points x_i du tableau équidistants: $x_i = x_0 + ih$, et soient f_i les valeurs correspondantes de la fonction. h est appelé *pas du tableau*.

Les différences $f_{i+1} - f_i$ sont dites *différences du premier ordre*. Selon la nature du problème, cette quantité se note soit Δf_i (*différences ascendantes*), soit ∇f_{i+1} (*différences descendantes*), soit encore $\delta f_{i+1/2} = f^1_{i+1/2}$ (*différences centrales*). Ainsi,

$$f_{i+1} - f_i = \Delta f_i = \nabla f_{i+1} = \delta f_{i+1/2} = f^1_{i+1/2}. \tag{1}$$

Les *différences d'ordre supérieur* sont formées à l'aide des relations de récurrence

$$\begin{aligned}
\Delta^m f_i &= \Delta(\Delta^{m-1} f_i) = \Delta^{m-1} f_{i+1} - \Delta^{m-1} f_i, \\
\nabla^m f_i &= \nabla(\nabla^{m-1} f_i) = \nabla^{m-1} f_i - \nabla^{m-1} f_{i-1}, \\
\delta^m f_i &= \delta(\delta^{m-1} f_i) = \delta^{m-1} f_{i+1/2} - \delta^{m-1} f_{i-1/2}, \\
f^m_i &= f^{m-1}_{i+1/2} - f^{m-1}_{i-1/2}.
\end{aligned}$$

Le tableau des différences a en général l'aspect suivant:

x	f	f^1	f^2	f^3	
x_0	f_0	$f^1_{1/2}$			⋮
x_1	f_1	$f^1_{3/2}$	f^2_1	$f^3_{3/2}$	⋮
x_2	f_2	$f^1_{5/2}$	f^2_2	$f^3_{5/2}$	⋮
x_3	f_3	$f^1_{7/2}$	f^2_3		
x_4	f_4				
⋮	⋮	⋮	⋮	⋮	

Certaines formules d'interpolation emploient de plus les moyennes arithmétiques de deux quantités successives d'une même colonne:

$$\mu f^m_i = (f^m_{i+1/2} + f^m_{i-1/2})/2, \quad m \text{ impair}$$

$$\mu f^m_{i+1/2} = (f^m_{i+1} + f^m_i)/2, \quad m \text{ pair}.$$

LEMME. *Les différences m-ièmes s'expriment par les valeurs de la fonction selon la formule*

$$f^m_i = \sum_{j=0}^{m} (-1)^j C^j_m f_{i+m/2-j} \tag{2}$$

(*i est entier pour m pair et demi-entier pour m impair*).

La démonstration se fait par récurrence. Quand $m = 1$, la relation (2) est vraie d'après (1). Supposons qu'elle l'est pour $m = l$. On a

$$f^{l+1}_i = f^l_{i+1/2} - f^l_{i-1/2} = \sum_{j=0}^{l} (-1)^j C^j_l f_{i+1/2+l/2-j} - \sum_{j=0}^{l} (-1)^j C^j_l f_{i-1/2+l/2-j}.$$

En regroupant les coefficients des mêmes f_k et en utilisant l'égalité

$$C^j_l + C^{j+1}_l = C^{j+1}_{l+1},$$

on obtient l'expression cherchée de f^{l+1}_i. Le lemme est démontré.

(2) implique que l'*opérateur différence finie est linéaire*. En particulier, elle donne

$$f_i^2 = f_{i+1} - 2f_i + f_{i-1},$$
$$f_i^3 = f_{i+3/2} - 3f_{i+1/2} + 3f_{i-1/2} - f_{i-3/2},$$
$$f_i^4 = f_{i+2} - 4f_{i+1} + 6f_i - 4f_{i-1} + f_{i-2},$$

LEMME. *Pour* $x_i \equiv x_0 + ih$, *on a*

$$f(x_i; \ldots; x_{i+m}) = \frac{f_{i+m/2}^m}{h^m \cdot m!}. \tag{3}$$

Raisonnons par récurrence. Si $m = 1$, on a

$$f(x_i; x_{i+1}) = \frac{f_{i+1} - f_i}{x_{i+1} - x_i} = \frac{f_{i+1/2}^1}{h}.$$

Supposant que (3) est juste quels que soient $m \leqslant l$, il vient

$$f(x_i; \ldots; x_{i+l+1}) = \frac{f(x_{i+1}; \ldots; x_{i+l+1}) - f(x_i; \ldots; x_{i+l})}{x_{i+l+1} - x_i} =$$
$$= \frac{f_{i+1+l/2}^l - f_{i+l/2}^l}{(h^l l!)(h(l+1))} = \frac{f_{i+(l+1)/2}^{l+1}}{h^{l+1}(l+1)!};$$

donc, (3) est également juste pour $m = l + 1$, ce qui démontre le lemme.

Selon (5.4) on a $f(x_i; \ldots; x_{i+m}) = \frac{f^{(m)}(\zeta)}{m!}$ avec $x_i \leqslant \zeta \leqslant x_{i+m}$. La comparaison avec (3) donne

$$\Delta^m f_i = \nabla^m f_{i+m} = \delta^m f_{i+m/2} = f_{i+m/2}^m = h^m f^{(m)}(\zeta). \tag{4}$$

CONSÉQUENCE. *Les différences finies d'ordre n d'un polynôme de degré n sont constantes et les différences d'ordre supérieur sont nulles.*

En calcul manuel, on profite souvent de cette dernière circonstance pour tabuler les polynômes. On calcule directement les valeurs du polynôme pour $n + 1$ valeurs consécutives de l'argument, puis on forme le tableau des différences jusqu'à l'ordre n inclus. On remplit ensuite la colonne des différences d'ordre n (compte tenu de leur constance), celle des différences d'ordre $n - 1$, puis des différences d'ordre $n - 2$, et ainsi de suite. Ce faisant, on utilise la formule

$$f_i^l = f_{i-1}^l + f_{i-1/2}^{l+1}.$$

On constate qu'en dressant le tableau de $N \gg n$ valeurs du polynôme de degré n, on commet, au pis aller, une erreur de calcul de l'ordre de $N^n \varepsilon$, ε étant l'unité de la dernière décimale utilisée. La méthode considérée exige donc une grande prudence. En particulier, lorsque le tableau est trop grand, on le divise en panneaux et on opère dans chaque panneau pris à part.

Examinons l'influence de l'erreur commise sur une valeur f_i quelconque sur les différences finies de divers ordres; supposons par exemple qu'on a $f_{i+\varepsilon}$ au lieu de f_i. Le tableau des différences a alors la forme

$$\begin{array}{lll}
f_{i-2} & & \\
 & f^1_{i-3/2} & \\
f_{i-1} & & f^2_{i-1}+\varepsilon \\
 & f^1_{i-1/2}+\varepsilon & \\
f_i+\varepsilon & & f^2_i-2\varepsilon \\
 & f^1_{i+1/2}-\varepsilon & \\
f_{i+1} & & f^2_{i+1}+\varepsilon \\
 & f^1_{i+3/2} & \\
f_{i+2} & &
\end{array}$$

Nous voyons qu'en conformité avec (2), les différences d'ordre m sont entachées d'une erreur de coefficients $(-1)^j C^j_m$. Si la fonction est suffisamment régulière, ses différences d'ordre pas très élevé peuvent s'avérer petites, et les quantités $C^j_m\varepsilon$ paraîtront en comparaison suffisamment grandes. Un simple examen du tableau des différences permet de corriger la valeur de la fonction qui contient une erreur.

On détecte de même les erreurs résultant de la tabulation des différences. Soit, par exemple,

$$f^3_{1/2}=1\cdot10^{-5}, \qquad f^3_{3/2}=2\cdot10^{-5}, \qquad f^3_{5/2}=12\cdot10^{-5},$$
$$f^3_{7/2}=-23\cdot10^{-5}, \qquad f^3_{9/2}=11\cdot10^{-5}, \qquad f^3_{11/2}=1\cdot10^{-5}.$$

Si une valeur f_i quelconque présentait une erreur ε relativement grande, cela entraînerait pour les différences finies troisièmes l'apparition des quantités ε, -3ε, 3ε, $-\varepsilon$. Dans le cas considéré les différences d'ordre trois sont en fait nulles, à l'exception de $f^3_{5/2}$, $f^3_{7/2}$, $f^3_{9/2}$, qui sont approximativement de la forme ε, -2ε, ε ($\varepsilon = 11\cdot10^{-5}$). Ce qui laisse à penser qu'une erreur a été commise dans le calcul de $f^1_{7/2}$, erreur qui a justement perturbé les différences troisièmes.

Signalons une propriété des tableaux de différences des fonctions. Puisque les valeurs de $f(x_i)$ ne sont pas obtenues avec une précision absolue, mais sont arrondies pendant et après les calculs, on a en fait, au lieu de f_i, les valeurs $y_i = f_i + \eta_i$. Les valeurs des différences s'écrivent alors

$$y^m_i = f^m_i + \eta^m_i.$$

Pour $f(x)$ suffisamment régulière (nous ne spécifierons pas la nature de ce phénomène), les valeurs f^m_i décroîtront avec la croissance de m. Ainsi, on a pour $f(x) = \sin x$

$$f^m_i = h^m f^{(m)}(\zeta), \qquad |f^{(m)}(\zeta)| \leqslant 1;$$

aussi $|f^m_i| \leqslant h^m \to 0$ avec $h < 1$, $m \to \infty$. Les quantités η^m_i s'écrivent comme des combinaisons de termes de la forme $(-1)^j C^j_m \eta_k$ et peuvent donc croître

avec m. C'est pourquoi, si l'on effectue des arrondis, la fonction $\varphi(m) = \max_i |y_i^m|$ décroît pour m petits et croît pour m grands. Dans le domaine des valeurs de m où les différences d'ordre supérieur commencent à croître, l'information la plus importante fournie par le tableau des différences d'ordre k d'une fonction régulière est celle sur les différences d'erreurs η_i^k, de sorte que généralement ces différences ne nous apportent aucune information sur la fonction.

§ 11. Formules d'interpolation de Newton pour des points équidistants

Pour approcher une fonction avec une précision élevée on opère de deux façons:

1. On ajoute de nouveaux points d'interpolation et la fonction est interpolée sur tous les points par un polynôme de plus haut degré possible.

2. Une fois les nouveaux points ajoutés, on interpole sur les points les plus proches; le degré du polynôme d'interpolation ne s'élève pas.

Considérons le deuxième procédé. Si le choix des points n'est pas spécifié, dans le cas d'un grand tableau des valeurs de la fonction, on prend généralement des points régulièrement espacés: $x_j = x_0 + jh$. Si les points sont choisis dans le voisinage du point x en lequel on calcule la fonction, le point intermédiaire ζ dans l'estimation du reste (3.1) est contenu lui aussi dans ce voisinage de sorte que $f^{(n)}(\zeta)$ ne varie pas très sensiblement pour un tel choix.

Par conséquent, l'erreur se ressentit surtout de la quantité

$$|\omega_n(x)| = \prod_{j=1}^{n} |x - x_j|,$$

c'est-à-dire du produit des distances du point x aux points d'interpolation. La valeur de $|\omega_n(x)|$ sera donc minimale si l'on prend, pour trouver $f(x)$, n points les plus proches de x. Il est évident que pour $n = 2l$ (pair) on choisit l points à gauche et l points à droite de x et, pour $n = 2l + 1$ (impair), le point le plus proche de x et l points de part et d'autre de cette valeur. Si x se trouve au début ou à la fin du tableau, ce choix est impossible.

Pour interpoler au début (resp. à la fin) du tableau, on représente le polynôme par les *formules de Newton* d'interpolation en avant (resp. en arrière). Soit $L_n(x)$ le polynôme de Lagrange basé sur $x_0, \ldots, x_{n-1}$. Selon (5.3) on a

$$L_n(x) = f(x_0) + f(x_0; x_1)(x - x_0) + \ldots + f(x_0; \ldots; x_{n-1})(x - x_0) \ldots (x - x_{n-2}).$$

Effectuons le changement de variables $x = x_0 + ht$ et passons des différences divisées aux différences finies d'après (10.3). Alors

$$L_n(x_0 + ht) = f_0 + f_{1/2}^1 t + \ldots + f_{(n-1)/2}^{n-1} \frac{t|(t-1)\ldots(t-(n-2))}{(n-1)!}. \quad (1)$$

Le reste (3.1) s'écrit

$$f^{(n)}(\zeta)\,\frac{t\,(t-1)\,\ldots\,(t-(n-1))\,h^n}{n\,!}\,.$$

La formule (1) est la *formule de Newton pour l'interpolation en avant.* Le même changement de variables dans le polynôme $L_n(x)$ basé sur $x_0,\ x_{-1},\ \ldots,\ x_{-(n-1)}$

$$L_n(x) = f(x_0) + f(x_0;\ x_{-1})\,(x-x_0) + \ldots$$
$$\ldots + f(x_0;\ \ldots;\ x_{-(n-1)})\,(x-x_0)\ \ldots\ (x-x_{-(n-2)})$$

donne la *formule de Newton pour l'interpolation en arrière*

$$L_n(x_0+ht) = f_0 + f^1_{-1/2}t + \ldots + f^{n-1}_{-(n-1)/2}\frac{t\,(t+1)\,\ldots\,(t+(n-2))}{(n-1)\,!} \qquad (2)$$

avec le reste

$$f^{(n)}(\zeta)\,\frac{t\,(t+1)\,\ldots\,(t+(n-1))\,h^n}{n\,!}\,.$$

Voyons maintenant comment évaluer pratiquement une erreur. Si la fonction $f(x)$ est donnée analytiquement, on peut estimer sa dérivée et en tirer l'évaluation du reste de l'interpolation. Pour une fonction assez compliquée, cette opération risque d'exiger beaucoup de temps et des exécutants suffisamment qualifiés; d'autre part, l'estimation obtenue s'avère très souvent trop forte. On sacrifie donc généralement la rigueur à la simplicité et on remplace la borne supérieure de la dérivée par une valeur approchée de cette borne.

S'il s'agit d'évaluer l'erreur $f(x) - L_n(x)$ en un point x concret, on a, conformément à (5.6)-(5.7),

$$f(x) - L_n(x) \approx f(x_1;\ \ldots;\ x_{n+1})\,\omega_n(x).$$

Dans d'autre cas, c'est la petitesse de l'erreur d'interpolation en tous les points du segment $[a,\ b]$ qui est essentielle, et il est indispensable d'évaluer

$$M_n = \max_{[a,\,b]} |f^{(n)}(x)|.$$

Pour fixer les idées, supposons qu'il faut interpoler une fonction donnée sur un réseau à pas constant: $x_q = x_0 + qh$. Si $f^{(n)}(x)$ est continue, il vient

$$\lim_{h\to 0}\sigma_n^h = M_n;\quad \sigma_n^h = \max_{a\leqslant x_m\leqslant b-nh}\left|\frac{\Delta^n f_m}{h^n}\right|,$$

pour h petits, on pose donc $M_n \approx \sigma_n^h$.

§ 12. Formules d'interpolation de Bessel et d'Everett. Formation des tableaux

Dans de nombreux cas, la petitesse du degré des polynômes approchant une fonction est primordiale. On arrive parfois à abaisser le degré de ces polynômes sans perte de précision si l'on forme leurs combinaisons linéaires. Considérons des procédés d'approximation de fonctions parmi les plus élémentaires.

Si la valeur de x pour laquelle on fait l'interpolation appartient à $(x_0, x_0 + h/2]$, on choisira les abscisses dans l'ordre suivant:

$$x_0, x_0 + h, x_0 - h, \ldots, x_0 + lh, x_0 - lh.$$

Alors, pour q quelconque, le groupe des q premiers termes de cette suite est formé des points les plus proches de x et l'estimation du reste

$$|f(x) - L_n(x)| \leqslant \max_x \frac{|f^{(n)}(x)|}{n!} |\omega_n(x)$$

pour le polynôme défini sur cet ensemble d'abscisses est plus précise que pour tout autre ensemble. La formule d'interpolation écrite pour les premiers points de la suite ci-dessus est la *formule de Gauss pour l'interpolation en avant*. La suite $x_0, x_0 - h, x_0 + h, \ldots$ sera la meilleure au même sens pour l'interpolation en un point de $[x_0 - h/2, x_0)$. La formule correspondante est dite *formule de Gauss pour l'interpolation en arrière*. Les formules de Gauss ont une valeur intrinsèque et s'appliquent largement en calcul manuel. Ecrivons la formule de Gauss pour l'interpolation en avant à $2n + 2$ points

$$L_{2n+2}(x) = f(x_0) + f(x_0; x_1)(x - x_0) + \ldots$$
$$\ldots + f(x_0; x_1; x_{-1}; \ldots; x_{n+1})(x - x_0) \ldots$$
$$\ldots (x - x_n)(x - x_{-n}). \quad (1)$$

La substitution $x = x_0 + ht$ fournit

$$G_{2n+2}(x_0 + ht) = L_{2n+2}(x_0 + ht) =$$
$$= f_0 + f^1_{1/2} t + f^2_0 \frac{t(t-1)}{2!} + \ldots + f^{2n+1}_{1/2} \frac{t(t^2-1)\ldots(t^2-n^2)}{(2n+1)!}. \quad (2)$$

Ecrivons maintenant la formule de Gauss du même ordre pour l'interpolation en arrière par rapport à x_1 (c'est-à-dire pour l'ensemble des points $x_1, x_0, x_2, x_{-1}, \ldots, x_{n+1}, x_{-n}$):

$$L_{2n+2}(x) = f(x_1) + f(x_1; x_0)(x - x_1) + \ldots$$
$$\ldots + f(x_1; x_0; \ldots; x_{n+1}; x_{-n})(x - x_1)(x - x_0) \ldots$$
$$\ldots (x - x_{n+1}). \quad (3)$$

On obtient par le changement de variables $x = x_0 + ht$

$$G_{2n+2}(x_0 + ht) = f_1 + f^1_{1/2}(t-1) + f^2_1 \frac{t(t-1)}{2!} + \ldots$$
$$\ldots + f^{2n+1}_{1/2} \frac{t(t^2-1)\ldots(t^2-(n-1)^2)(t-n)(t-(n+1))}{(2n+1)!}. \quad (4)$$

Les formules (1), (2), (3) et (4) sont en fait les diverses écritures d'un même polynôme de Lagrange basé sur $x_{-n}, \ldots, x_{n+1}$. On peut donc dire que ce polynôme est égal à la demi-somme des seconds

membres de (2) et (4). Notons à ce propos que les différences d'ordre impair des deux formules sont identiques et qu'on peut, par conséquent, réduire les termes semblables correspondant aux mêmes différences. Par ailleurs, les coefficients des différences d'ordre pair sont égaux et peuvent être regroupés. On aboutit à la *formule d'interpolation de Bessel*

$$B_{2n+2}(x_0+ht)=G_{2n+2}(x_0+ht)=L_{2n+2}(x_0+ht)=$$
$$=\mu f_{1/2}+f^1_{1/2}(t-1/2)+\mu f^2_{1/2}\frac{t(t-1)}{2!}+\dots$$
$$\dots+\mu f^{2n}_{1/2}\frac{t(t^2-1)\dots(t^2-(n-1)^2)(t-n)}{(2n)!}+$$
$$+f^{2n+1}_{1/2}\frac{t(t^2-1)\dots(t^2-(n-1)^2)(t-n)(t-1/2)}{(2n+1)!} \qquad (5)$$

(ici et dans la suite on omet les calculs intermédiaires). Le reste de cette formule coïncide naturellement avec le reste du polynôme d'interpolation de Lagrange

$$f(x)-B_{2n+2}(x)=f^{(2n+2)}(\zeta)\frac{(x-x_{-n})\dots(x-x_n)(x-x_{n+1})}{(2n+2)!}$$

ou

$$f(x_0+ht)-B_{2n+2}(x_0+ht)=$$
$$=f^{(2n+2)}(\zeta)\,h^{2n+2}\frac{t(t^2-1)\dots(t^2-n^2)(t-(n+1))}{(2n+2)!}. \qquad (6)$$

Le seul avantage de (5) par rapport aux formules de Lagrange et aux formules de Gauss (1) et (2) est d'être d'un emploi plus commode. En particulier, pour $t = 1/2$, i.e. pour le calcul par dichotomie, les coefficients des différences d'ordre impair s'annulent. Cependant, en rejetant le dernier terme du second membre de (5) on obtient

$$B_{2n+1}(x_0+ht)=\mu f_{1/2}+f^1_{1/2}(t-1/2)+$$
$$+\mu f^2_{1/2}\frac{t(t-1)}{2!}+\dots+\mu f^{2n}_{1/2}\frac{t(t^2-1)\dots(t^2-(n-1)^2)(t-n)}{(2n)!}, \qquad (7)$$

où le polynôme $B_{2n+1}(x_0 + ht)$ de degré $2n$ n'est plus un polynôme d'interpolation proprement dit. En effet, il s'obtient comme la demi-somme de polynômes d'interpolation aux abscisses

$$x_{-n}, \dots, x_n \quad \text{et} \quad x_{-(n-1)}, \dots, x_{n+1}$$

respectivement, c'est-à-dire qu'il ne coïncide avec $f(x)$ qu'aux $2n$ points $x_{-(n-1)}, \dots, x_n$, tandis que son expression utilise bien les valeurs de $f(x)$ en x_{-n} et x_{n+1} (fig. 2.12.1). Son reste est donc formé

du terme négligé du second membre de (5) et du second membre de (6):

$$f(x_0+ht)-B_{2n+1}(x_0+ht)=$$
$$=f^{2n+1}_{1/2}\frac{t(t^2-1)\dots(t^2-(n-1)^2)(t-n)(t-1/2)}{(2n+1)!}+$$
$$+f^{(2n+2)}(\zeta)\,h^{2n+2}\frac{t(t^2-1)\dots(t^2-n^2)(t-(n+1))}{(2n+2)!}. \qquad (8)$$

On se convainc que le premier terme à droite dans (8), qui est le terme principal, est inférieur à la partie principale de l'erreur

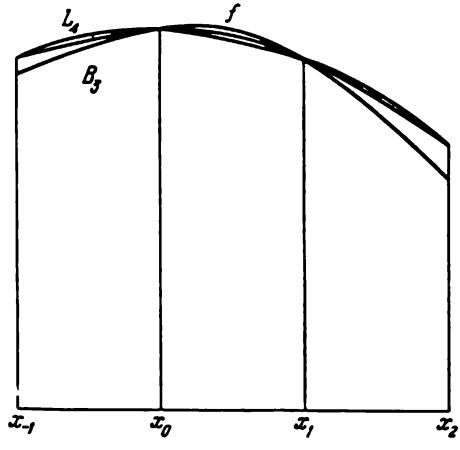


Fig. 2.12.1

dans chacune des formules dont la moyenne est notre polynôme $B_{2n+1}(x_0+ht)$. Dans la suite, nous vérifierons la dernière proposition pour $n=1$.

Prenons le problème type suivant. On demande de dresser le tableau de valeurs d'une fonction de façon que l'erreur due à l'interpolation de ces valeurs par un polynôme de degré m donné ne dépasse pas ε. On dit alors que le tableau permet une interpolation d'ordre m (avec une erreur ε). Les tableaux d'usage général autorisent d'habitude une *interpolation linéaire*. Notre exposé ultérieur utilisera les tableaux à pas constant. Pour calculer une valeur de $f(x)$ on choisit alors à droite et à gauche d'une valeur donnée x les points x_q et x_{q+1} respectivement (ce sont les points les plus proches de x). On remplace $f(x)$ par un polynôme du premier degré pour ces points:

$$f(x_q+th)\approx L_2(x_q+th)=f_q+f^1_{q+1/2}t.$$

L'erreur de cette formule est

$$f''(\zeta)\,h^2\frac{t(t-1)}{2},$$

où

$$x_q \leqslant \zeta \leqslant x_{q+1}.$$

Elle n'est pas supérieure à ε si

$$(\max |f''(\zeta)|)\, h^2 \max_{[0,\,1]} \left| \frac{t(t-1)}{2} \right| \leqslant \varepsilon;$$

$\max_{[0,1]} \left| \frac{t(t-1)}{2} \right|$ est atteint pour $t = 1/2$ et vaut 1/8. Il suffit donc que la condition

$$\max |f''(\zeta)|\, h^2/8 \leqslant \varepsilon \tag{9}$$

soit vérifiée.

Supposons que nous voulons établir ou enregistrer en machine un tableau de $\sin x$ sur $[0, \pi/2]$ tel que l'erreur d'interpolation linéaire soit au plus $0{,}5 \cdot 10^{-6}$. Vu que $\max |(\sin x)''| = 1$, il résulte de (9) que le pas doit être tel que

$$h^2/8 \leqslant 0{,}5 \cdot 10^{-6} \quad \text{ou} \quad h \leqslant 0{,}002.$$

La condition d'admissibilité de l'interpolation linéaire est quelquefois trop forte si bien qu'on lui préfère l'*interpolation quadratique* (interpolation au moyen d'un polynôme du second degré). Un cas élémentaire de l'interpolation quadratique est l'interpolation par le polynôme de Lagrange sur les trois points les plus proches. Soit x_q le point le plus proche de x. Posons $x = x_q + ht$ et

$$f(x_q - ht) \approx L_3(x_q + ht) = f_q + f^1_{q+1/2} t + f^2_q \frac{t(t-1)}{2}. \tag{10}$$

Le reste de (10) est

$$f^{(3)}(\zeta)\, h^3 \frac{t(t^2-1)}{3!}.$$

Pour qu'un tableau donné se prête à l'interpolation quadratique (10) il suffit d'avoir

$$\max |f^{(3)}(\zeta)|\, h^3 \max_{|t| \leqslant 1/2} \left| \frac{t(t^2-1)}{6} \right| \leqslant \varepsilon.$$

Puisque $\max_{|t| \leqslant 1/2} \left| \frac{t(t^2-1)}{6} \right| = 1/16$, cette condition se récrit pour un seul pas:

$$\max |f^{(3)}(\zeta)|\, h^3/16 \leqslant \varepsilon. \tag{11}$$

Dans le cas concret $f(x) = \sin x$, $\varepsilon = 0{,}5 \cdot 10^{-6}$, on obtient $h \leqslant 0{,}02$.

Voici une autre technique permettant de remplacer la fonction par un polynôme de degré 2. Soit $x \in (x_q, x_{q+1})$, $x = x_q + ht$. On pose

$$f(x_q + ht) \approx B_3(x_q + ht) =$$

$$= \mu f_{q+1/2} + f^1_{q+1/2}(t - 1/2) + \mu f^2_{q+1/2} \frac{t(t-1)}{2}, \tag{12}$$

autrement dit, on remplace $f(x)$ par le polynôme de Bessel du second degré (7) défini pour les points x_{q-1}, x_q, x_{q+1}, x_{q+2}. Selon (8), le reste de (12) est

$$f^3_{q+1/2}\frac{t(t-1)(t-1/2)}{6}+f^{(4)}(\zeta)h^4\frac{t(t^2-1)(t-2)}{24}.$$

Du moment que $f^3_{q+1/2}=h^3f^{(3)}(\tilde{\zeta})$, l'interpolation par la formule (12) se fera si

$$\max|f^{(3)}(\zeta)h^3\max_{0\leqslant t\leqslant 1}\left|\frac{t(t-1)(t-1/2)}{6}\right|+$$

$$+\max|f^{(4)}(\zeta)|h^4\max_{0\leqslant t\leqslant 1}\left|\frac{t(t^2-1)(t-2)}{24}\right|\leqslant\varepsilon.$$

Puisque

$$\max_{0\leqslant t\leqslant 1}\left|\frac{t(t-1)(t-1/2)}{6}\right|=\frac{1}{72\sqrt{3}},\quad\max_{0\leqslant t\leqslant 1}\left|\frac{t(t^2-1)(t-2)}{24}\right|=\frac{3}{128},$$

elle l'est pour

$$\frac{\max|f^{(3)}(\zeta)|h^3}{72\sqrt{3}}+\frac{3\max|f^{(4)}(\zeta)|}{128}h^4\leqslant\varepsilon. \tag{13}$$

Pour h petits, le premier terme constitue la partie principale; il est $9\sqrt{3}/2=7{,}794\ldots$ fois inférieur au premier membre de (11). Par conséquent, h étant petit, on vérifie (13) avec un pas $\sqrt[3]{9\sqrt{3}/2}\approx$ $\approx 1{,}98$ fois supérieur à celui du cas (11). Dans l'exemple considéré la condition (13) s'écrit

$$h^3/72\sqrt{3}+3h^4/128\leqslant 0{,}5\cdot 10^{-6}.$$

En la résolvant, on trouve $h\leqslant h_0=0{,}038\ldots$ La valeur 0,038 du pas n'est pas commode en calcul à la main si bien qu'en dressant le tableau correspondant, on l'aurait probablement remplacée par une valeur arrondie, à savoir 0,03.

Si le pas est néanmoins grand, on augmente le degré du polynôme d'interpolation.

Et voici un autre procédé d'écriture du polynôme d'interpolation dit *méthode d'Everett*, qui se prête parfaitement à la résolution du problème suivant: soit un tableau de valeurs d'une fonction à points x_0+nh; on demande de former le tableau à points x_0+nh' pour $h'=h/l$, l entier. Il est douteux que les constructions ci-dessous liées aux formules d'Everett puissent être utiles au lecteur. Elles présentent cependant un certain intérêt en tant que procédés particuliers et problèmes variationnels particuliers permettant d'augmenter l'efficacité des algorithmes de calcul dans telle situation concrète.

Revenons à la formule de Gauss (2) et éliminons-en les différences d'ordre impair par la relation

$$f^{2l+1}_{1/2}=f^{2l}_1-f^{2l}_0.$$

Après réduction des termes semblables on débouche sur la formule d'Everett

$$E_{2n+2}(x_0 + ht) = G_{2n+2}(x_0 + ht) = S_0(u) + S_1(t); \tag{14}$$

ici $u = 1 - t$

$$S_q(t) = f_q t + f_q^2 \frac{t(t^2-1)}{3!} + \ldots + f_q^{2n} \frac{t(t^2-1)\ldots(t^2-n^2)}{(2n+1)!}. \tag{15}$$

Soulignons une fois de plus que toutes ces formules représentent un même polynôme d'interpolation. Le reste de cette formule est celui de la formule de Lagrange, i.e.

$$f^{(2n+2)}(\zeta)\, h^{2n+2} \frac{t(t^2-1)\ldots(t^2-n^2)(t-(n+1))}{(2n+2)!}. \tag{16}$$

L'écriture (14) du polynôme d'interpolation a l'avantage de réduire de moitié le travail nécessité par le remplissage des tableaux: dans le cas de la formule d'Everett, les valeurs $f(x_0 - ht)$, $0 < t < 1$, se calculent par

$$f(x_0 - ht) = S_0(u) + S_{-1}(t);$$

ainsi, la valeur $S_0(u)$ sert à trouver $f(x_0 \pm ht)$. On peut calculer tous les $S_q(u)$ pour $u = 1/l, \ldots, (l-1)/l$ et former les valeurs de $f(x)$ aux nœuds $x_0 + nh'$ du réseau serré comme combinaisons linéaires. Le calcul direct du polynôme d'interpolation à l'aide des autres écritures veut qu'on additionne $2n + 2$ termes contre $n + 1$ termes pour (15).

Effectuons quelques constructions supplémentaires pour $n = 2$ lorsque

$$S_q(t) = f_q t + f_q^2 \frac{t(t^2-1)}{6} + f_q^4 \frac{(t^2-1)(t^2-4)}{120}.$$

En calcul manuel il est utile d'approcher le coefficient de f_q^4 par $-k\dfrac{t(t^2-1)}{120}$ et de calculer

$$\overline{S}_q(t) = f_q t + \left(f_q^2 - \frac{k}{20} f_q^4\right) \frac{t(t^2-1)}{6}$$

au lieu de $S_q(t)$. On peut offrir aux praticiens du calcul numérique des tableaux qui contiennent, à côté de la colonne des f_q, celle des $f_q^2 - \dfrac{k}{20} f_q^4$. Le calcul des valeurs de la fonction par la formule

$$f(x_q + ht) \approx \overline{S}_{q+1}(t) + \overline{S}_q(u) = \overline{E}_6(x_q + ht)$$

serait alors des plus faciles. Proposons-nous de choisir le plus raisonnablement le paramètre k. Nous énoncerons ce problème comme celui du choix d'un k tel que la partie principale

$$\sup_{t \in [0,\,1]} |E_6(x_q + ht) - \overline{E}_6(x_q + ht)|$$

soit minimale. On a

$$E_6(x_q + ht) - \overline{E}_6(x_q + ht) = f_{q+1}^4 \frac{t(t^2-1)(t^2-4+k)}{120} + f_q^4 \frac{u(u^2-1)(u^2-4+k)}{120}. \tag{17}$$

Il paraît de prime abord que le mieux serait de choisir k à partir de la condition

$$\min_k \max_{[0,\,1]} |\, t\,(t^2-1)\,(t^2-4+k)\,|.$$

f^4_{q+1} et f^4_q étant voisins, les termes du second membre de (17) peuvent se compenser, et ce choix n'est donc pas le meilleur. Isolons donc au préalable le terme principal à droite dans (17) et portons

$$f^4_{q+1}=\mu f^4_{q+1/2}+1/2 f^5_{q+1/2},\quad f^4_q=\mu f^4_{q+1/2}-1/2 f^5_{q+1/2},\quad u=1-t.$$

En réduisant les termes semblables, il vient

$$E_6(x_q+ht)-\overline{E}_6(x_q+ht)=s_1+s_2,$$

où

$$s_1=\mu f^4_{q+1/2}\,\frac{5t\,(t^2-1)\,(t-2)+3kt\,(t-1)}{120}, \tag{18}$$

$$s_2=f^5_{q+1/2}\,\frac{t\,(t^2-1)\,(t-2)\,(t-1/2)+kt\,(t-1)\,(t-1/2)}{120}. \tag{19}$$

Si $f^{(4)}(x)\neq 0$, s_1 est de l'ordre de h^4. C'est donc la partie principale de la différence $E_6(x_q+ht)-\overline{E}_6(x_q+ht)$. On choisit le plus naturellement k à partir de la condition de minimum du coefficient de l'expression de s_1:

$$\min_k \max_{[0,\,1]} |\, 5t\,(t^2-1)\,(t-2)+3kt\,(t-1)\,|.$$

On a l'égalité

$$t\,(t^2-1)\,(t-2)=t\,(t-1)\,(t\,(t-1)-2).$$

Après avoir effectué le changement de variable $y=4t\,(1-t)$, on obtient, par conséquent,

$$5t\,(t^2-1)\,(t-2)+3kt\,(t-1)=\frac{5}{16}\,y^2+\frac{5}{2}\,y-\frac{3}{4}\,ky.$$

Puisque pour t variant sur l'intervalle fermé $[0,\,1]$, y parcourt deux fois cet intervalle, on a

$$\min_k \max_{[0,\,1]} |\, 5t\,(t^2-1)\,(t-2)+3kt\,(t-1)|=$$

$$=\min_k \max_{[0,\,1]}\left|\,\frac{5}{16}\,y^2+\frac{5}{2}\,y-\frac{3}{4}\,ky\,\right|=\frac{5}{16}\min_r \max_{[0,\,1]} |\, y^2-ry\,|, \tag{20}$$

avec $r=\dfrac{12}{5}\,k-8$.

Examinons la figure 2.12.2. Pour $0<r<1$, la fonction y^2-ry admet sur $[0,\,1]$ des extréma locaux en 1 et au point en lequel $(y^2-ry)'=0$, i.e. pour $y=r/2$. Le premier extrémum vaut $1-r$, le second $-r^2/4$. Définissons $r_0\in(0,\,1)$ par la condition que les valeurs de la fonction y^2-ry en ces points d'extréma locaux sont égales en valeur absolue, c'est-à-dire par l'équation $1-r_0=r_0^2/4$. D'où $r_0=\sqrt{8}-2$ et

$$\max_{[0,\,1]} |\, y^2-r_0 y\,|=1-r_0=3-\sqrt{8}=\frac{1}{3+\sqrt{8}}=b.$$

Si $r=r_1<r_0$, alors $\max\limits_{[0,\,1]}|y^2-ry|\geqslant 1-r>1-r_0=b$; si $r=r_2>r_0$, alors $\max\limits_{[0,\,1]}|y^2-ry|\geqslant\left|\left(\frac{r}{2}\right)^2-\frac{r^2}{2}\right|=\frac{r^2}{4}>\frac{r_0^2}{4}=1-r_0=b$; la validité de ces relations ressort directement de la figure examinée. Nous avons trouvé un r

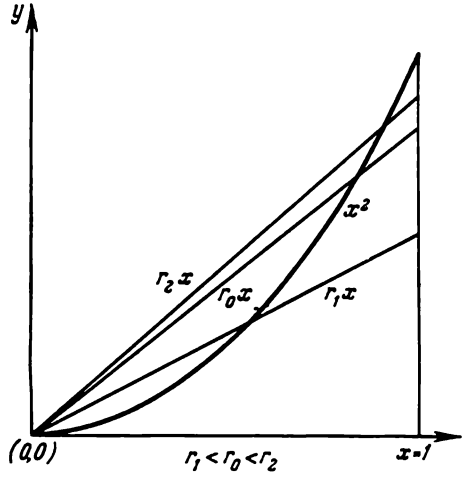


Fig. 2.12.2

tel que

$$\min_{r}\max_{[0,\,1]}|y^2-ry|=\frac{1}{3+\sqrt{8}}.$$

Compte tenu de (20), on obtient pour k:

$$k=5\,(3+\sqrt{2})/6=3{,}6785\ldots$$

et alors

$$\max_{[0,\,1]}|5t\,(t^2-1)\,(t-2)+3kt\,(t-1)=\frac{5}{16\,(3+\sqrt{8})}.$$

(18) permet l'estimation

$$|s_1|\leqslant\frac{\max|f^{(4)}(x)|}{384\,(3+\sqrt{8})}\,h^4=\frac{\max|f^{(4)}(x)|}{2238{,}0\ldots}\,h^4.$$

Evaluons $|s_2|$ pour ce k:

$$|s_2|\leqslant\frac{\max|f^{(5)}(x)|}{1646,\ldots}\,h^5.$$

D'autre part, on a pour le reste de départ

$$\frac{\max|f^{(6)}(x)}{204{,}8}\,h^6.$$

En additionnant, il vient

$$|f(x_q+ht)-\overline{E}_6(x_q+ht)| \leqslant \frac{\max|f^{(4)}(x)|}{2238{,}0\ldots}h^4+$$
$$+\frac{\max|f^{(5)}(x)|}{1646,\ldots}h^5+\frac{\max|f^{(6)}(x)|}{204{,}8}h^6.$$

On en tire que pour notre exemple $f(x)=\sin x$ l'erreur d'interpolation dans cette formule ne dépasse pas $0{,}5\cdot10^{-6}$ pour $h \leqslant h_0 = 0{,}16\ldots$ Ainsi, par rapport à l'interpolation linéaire, nous sommes arrivés à augmenter sensiblement le pas des tableaux pour une précision de l'interpolation de $0{,}5\cdot10^{-6}$.

Il se peut que la forme même de l'approximation cherchée d'une somme contenant des différences quatrièmes soit sans importance, ce qui a, par exemple, lieu si l'on se propose d'étudier les propriétés d'une fonction, de rechercher ses extréma, etc. Pour le faire, on essayera d'approcher ces termes par l'expression $f^4_{q+1}Q_3(t)+f^4_qQ_3(u)$ de façon que le maximum de la partie principale de l'erreur soit le plus petit possible.

PROBLÈME. Montrer que la solution du problème susmentionné est

$$Q_3(t)=\frac{960t(1-t^2)-5}{30720}$$

et que la partie principale de l'erreur ne dépasse pas

$$\frac{\max|f^{(4)}(x)|}{3072}h^4.$$

§ 13. Erreur d'arrondi dans l'interpolation

Supposons que nous sommes fixés sur le procédé d'interpolation à employer. L'exposé antérieur nous a donné une idée assez complète de l'erreur résultant de la substitution d'un polynôme à la fonction. Or, ce n'est pas la seule. En effet, les valeurs données de la fonction sont affectées d'erreurs dues, entre autres, à l'arrondi. Calculons la valeur de $f(x)$ par la formule

$$f(x)\approx\sum_{j=1}^{n}f_jP_j(x),$$

qui représente la forme générale des formules d'interpolation considérées. Puisqu'on se donne en fait non pas f_j, mais leurs valeurs approchées $\tilde{f}_j=f_j+\eta_j$, on obtient en définitive

$$f^*=\sum_{j=1}^{n}f_jP_j(x)+\sum_{j=1}^{n}\eta_jP_j(x).$$

Connaissant les limites entre lesquelles varie η_j, on peut évaluer la borne supérieure de l'erreur

$$\varepsilon=\sum_{j=1}^{n}\eta_jP_j(x).$$

Ainsi, pour $|\eta_j| \leqslant \eta$, on a

$$|\varepsilon| \leqslant \eta \sum_{j=1}^{n} |P_j(x)|.$$

On établit à l'usage général des tableaux à p décimales tels que l'interpolation linéaire avec une erreur de $0{,}5 \cdot 10^{-p}$ soit possible et que toutes les valeurs de la fonction soient fausses de $0{,}5 \cdot 10^{-p}$ au plus. Si l'on calcule une valeur de $f(x)$ au moyen d'une interpolation sur les points x_0, x_1 $(x_0 < x < x_1)$, alors

$$P_0(x) = \frac{x_1 - x}{x_1 - x_0}, \qquad P_1(x) = \frac{x - x_0}{x_1 - x_0}$$

et

$$\eta\,(|P_0(x)| + |P_1(x)|) = \eta.$$

On voit qu'en arrondissant les valeurs de la fonction l'erreur d'interpolation linéaire ne dépasse pas l'erreur sur ces valeurs. Etant donné que la première n'est pas supérieure à $0{,}5 \cdot 10^{-p}$, la deuxième erreur est de 10^{-p} au plus à condition que l'interpolation n'introduise aucune erreur supplémentaire.

La formation des tableaux à p décimales dans lesquels les valeurs de la fonction sont entachées d'erreurs inférieures ou égales à $0{,}5 \cdot 10^{-p}$ est une opération fort pénible. Soit

$$0,\ y_1 \ldots y_p \underbrace{499 \ldots 9}_{k \text{ fois}} \tag{1}$$

la valeur exacte de $f(x_1)$ ($y_1, \ldots, y_p$ décimales). On doit prendre pour f_1 la valeur $0, y_1, \ldots y_p$. Or, si, en calculant $f(x_1)$, on commet une erreur de $2 \cdot 10^{-k-p-1}$, la valeur choisie de f_1 sera $0, y_1 \ldots y_p + 10^{-p}$, soit une erreur $> 0{,}5 \cdot 10^{-p}$. Puisque k peut être aussi grand que l'on veut, on essayera de calculer $f(x_1)$ en travaillant sur un très grand nombre de chiffres. Dans la plupart des cas ou bien les tableaux à p décimales contiennent les valeurs de la fonction qui ne sont en erreur que de 10^{-p} au plus, ou bien on spécifie que l'erreur ne dépasse pas $0{,}5 \cdot 10^{-p}$ encore qu'elle soit en réalité de $0{,}5 \cdot 10^{-p}(1 + \gamma)$ avec γ suffisamment petit. L'erreur sur les valeurs de la fonction résulte également de l'erreur d'arrondi dans les formules d'interpolation, qui dépend naturellement du nombre de chiffres significatifs sur lesquels on opère. Dans ce cas, le nombre fixe de chiffres ne garantit pas non plus une erreur d'arrondi inférieure à $0{,}5 \times \times 10^{-p}$. En effet, si le polynôme d'interpolation possède la valeur (1) et si l'on arrondit ensuite à $2 \cdot 10^{-k-p-1}$ près, l'erreur sur la valeur ainsi obtenue peut être légèrement supérieure à $0{,}5 \cdot 10^{-p}$. Ainsi, même dans l'interpolation du tableau à p décimales vérifiant des conditions sévères imposées, les valeurs de la fonction arrondies à p décimales peuvent être en erreur de plus de $1{,}5 \cdot 10^{-p}$.

L'implantation des ordinateurs et une typographie automatique des tableaux réduisent la portée des questions que nous venons d'aborder. Les vieux traités de calcul numérique contiennent incomparablement plus de formes d'écriture de polynômes d'interpolation que les ouvrages actuels. C'est que dans le cas de la plupart des machines à calculer électroniques modernes, on travaille sur

des nombres de grande longueur en temps fixe indépendant du nombre de chiffres significatifs. La probabilité d'une panne ne dépend pas non plus de ce nombre. Aussi, les erreurs de méthode et d'arrondi restant les mêmes, la qualité de la formule d'interpolation n'est-elle fonction que du nombre d'opérations arithmétiques et logiques qu'elle nécessite. En calcul manuel le temps de calcul et la probabilité d'une panne (i.e. d'une erreur de l'exécutant) augmentent avec le nombre de chiffres significatifs utilisés. Les nombreuses écritures d'une même formule d'interpolation de Lagrange s'expliquaient pour beaucoup par le désir d'obtenir des formules employant les nombres de longueur minimale.

§ 14. Applications de l'interpolation. Interpolation inverse

On utilisera avec profit les polynômes d'interpolation dans certains problèmes.

Soit à trouver, par exemple, la valeur extrémale de la fonction d'une variable et le point d'extrémum correspondant. Dressons un tableau à grand pas de valeurs de la fonction en question. Son examen nous renseigne sur la position de l'extrémum. Sous-tabulons dans cette zone en utilisant un pas plus petit. Ce deuxième tableau précise la position de l'extrémum, et ainsi de suite. On se garde a de poursuivre indéfiniment ce remplissage du tableau. D'habitude, on désire connaître la valeur de l'extrémum avec une précision γ. Pour un certain pas le tableau permet une interpolation cubique, par exemple à l'aide des formules du § 12, avec la même précision γ. Remplaçons la fonction par un polynôme, dérivons celui-ci et trouvons un zéro. La valeur de la fonction au point correspondant est supposée valeur extrémale. Puisque $|f(x) - P_3(x)| \leqslant \gamma$, on a $|\max f(x) - \max P_3(x)| \leqslant \gamma$. Selon la nature du problème, on fractionnera le pas jusqu'à ce que l'interpolation quadratique devienne possible.

Un autre problème type se prêtant à l'interpolation est la recherche d'une racine X de l'équation $f(x) = d$.

La disposition des calculs est la même. On forme le tableau de la fonction pour définir *grosso modo* l'emplacement de la racine cherchée, on sous-tabule, et ainsi de suite.

Si le calcul de la fonction est relativement facile, on évitera d'interpoler à l'ordre supérieur à 2, sinon on se trouvera en présence d'un autre problème, celui de la recherche des zéros des polynômes, qui exige à lui seul une masse importante d'opérations arithmétiques. En cas d'un calcul pénible, on essayera d'élever le degré du polynôme d'interpolation.

Si, au voisinage de $y = d$, la fonction $g(y)$ inverse de $f(x)$ est suffisamment régulière, un procédé plus efficace sera l'*interpolation inverse*. Il s'agit de l'algorithme suivant. Supposons connues les valeurs $y_i = f(x_i)$ pour $i = 1, \ldots, n$, ce qui équivaut à dire qu'on connaît les valeurs $x_i = g(y_i)$ de la fonction inverse. Si l'interpolation en y est possible, on remplace $g(y)$ par le polynôme d'inter

polation $L_n(y)$ tel que

$$L_n(y_i) = x_i, \quad i = 1, \ldots, n,$$

et on pose $X = g(d) \approx L_n(d)$. Ce procédé est particulièrement indiqué si on veut ou bien connaître les valeurs de la solution de l'équation pour d en nombre suffisamment grand, ou bien obtenir explicitement la racine de $f(x) = d$ en fonction du paramètre d. Si l'interpolation avec les points $y_1, \ldots, y_n$ ne garantit pas la précision désirée, on fait $x_{n+1} = L_n(d)$, $y_{n+1} = f(x_{n+1})$; puis, selon le contexte, on remplace $g(d)$ par la valeur du polynôme d'interpolation pour tous les $y_1, \ldots, y_{n+1}$ ou pour les points les plus proches de d.

§ 15. Systèmes orthogonaux et leurs propriétés

Soit H un espace de Hilbert muni du produit scalaire (f, g) et, respectivement, de la norme $\| f \| = \sqrt{(f, f)}$.

Les conclusions les plus importantes du présent paragraphe sont vraies pour les cas suivants:

A) H est un espace de fonctions complexes définies sur $[a, b]$ avec l'intégrale bornée

$$\int_a^b p(x) |f(x)|^2 dx,$$

le produit scalaire est donné par l'égalité

$$(f, g) = \int_a^b p(x) f(x) \bar{g}(x) dx; \tag{1}$$

$p > 0$ presque partout sur $[a, b]$; les fonctions distinctes sur un ensemble de mesure nulle sont considérées comme égales.

B) H est un espace euclidien de vecteurs complexes n-dimensionnels $(x_1, \ldots, x_n)$ avec pour produit scalaire

$$(x, y) = \sum_{j=1}^{n} p_{ij} x_i \bar{y}_j, \tag{2}$$

$P = [p_{ij}]$ est une matrice définie positive.

Soit $\{\varphi_j\}_n = \varphi_1, \ldots, \varphi_n$ un système d'éléments de H. Ce système est dit *orthogonal* si $(\varphi_i, \varphi_j) = 0$ pour $i \neq j$ et *linéairement indépendant* si

$$\sum_{j=1}^{n} C_j \varphi_j = 0$$

à condition que tous les $C_j = 0$.

L'orthogonalisation d'un système donné d'éléments d'un espace hilbertien joue un grand rôle dans les recherches théoriques.

LEMME 1. *Soit un système linéairement indépendant d'éléments* $\varphi_1, \ldots, \varphi_n$. *On peut construire un système orthogonal linéairement indépendant*

$$\psi_j = \sum_{i=1}^{j} b_{ji}\varphi_i, \quad j = 1, \quad ., n, \tag{3}$$

où tous les $b_{jj} = 1$.

On se servira de la méthode de récurrence. Pour $n = 1$, on a le système trivial: $\psi_1 = \varphi_1$. Supposons que le système cherché $\{\psi_j\}_n$ soit construit pour un certain $n = k$. On trouve alors ψ_{k+1} sous la forme

$$\psi_{k+1} = \varphi_{k+1} - \sum_{i=1}^{k} a_{ki}\psi_i. \tag{4}$$

Les coefficients a_{ki} sont choisis à partir de la condition d'orthogonalité $(\psi_{k+1}, \psi_i) = 0$ avec $i \leqslant k$. Le système $\{\psi_j\}_k$ étant orthogonal, la dernière relation se récrit

$$(\varphi_{k+1}, \psi_i) - a_{ki}(\psi_i, \psi_i) = 0,$$

d'où

$$a_{ki} = \frac{(\varphi_{k+1}, \psi_i)}{(\psi_i, \psi_i)},$$

donc

$$\psi_{k+1} = \varphi_{k+1} - \sum_{i=1}^{k} \frac{(\varphi_{k+1}, \psi_i)}{(\psi_i, \psi_i)}\psi_i \tag{5}$$

est orthogonal par rapport à tous les éléments précédents. En y portant

$$\psi_i = \sum_{p=1}^{i} b_{ip}\varphi_p$$

($i \leqslant k$), on obtient la relation cherchée. L'ensemble de relations (3) lorsque $j \leqslant n$ peut être représenté par

$$\psi_{(n)} = B_n\varphi_{(n)},$$

où

$$\psi_{(n)} = \begin{pmatrix} \psi_1 \\ \vdots \\ \psi_n \end{pmatrix}, \quad \varphi_{(n)} = \begin{pmatrix} \varphi_1 \\ \vdots \\ \varphi_n \end{pmatrix}, \quad B_n = \begin{pmatrix} 1 & 0 & 0 & 0 & \ldots \\ b_{21} & 1 & 0 & 0 & \ldots \\ b_{31} & b_{32} & 1 & 0 & \ldots \\ \ldots & \ldots & \ldots & \ldots & \ldots \end{pmatrix}.$$

D'autre part, en transférant tous les ψ_l du second membre de (5) dans le premier, il vient

$$\varphi_{(n)} = A_n \psi_{(n)}, \quad \text{où} \quad A_n = \begin{pmatrix} 1 & 0 & 0 & 0 & \dots \\ a_{21} & 1 & 0 & 0 & \dots \\ a_{31} & a_{32} & 1 & 0 & \dots \\ \dots & \dots & \dots & \dots & \dots \end{pmatrix}. \tag{6}$$

La matrice B_n est quelquefois appelée *matrice d'orthogonalisation*. Puisque $\det B_n = 1$, la transformation définie par B_n est non dégénérée et change le système linéairement indépendant $\{\varphi_j\}_n$ en $\{\psi_j\}_n$ linéairement indépendant lui aussi.

Du moment que $\{\varphi_j\}_n$ est linéairement indépendant, il en découle $B_n = A_n^{-1}$.

Les systèmes de polynômes orthogonaux constituent un outil puissant de l'analyse classique. On prend en qualité du système de fonctions $\{\varphi_j\}$ le système $1, \dots, x^{j+1}, \dots$, qui est ensuite orthogonalisé dans l'espace doté du produit scalaire (1) selon le schéma décrit. Les polynômes

$$\psi_j(x) = x^{j-1} + \sum_{i=1}^{j-1} b_{ji} x^{i-1}, \tag{7}$$

que l'on obtient de cette façon, s'appellent *polynômes orthogonaux* par rapport au poids $p(x)$ et à l'intervalle $[a, b]$. Le nom de polynômes orthogonaux relativement au poids $p(x)$ est parfois donné à $\Psi_j(x) = \alpha_j \psi_j(x)$, α_j étant choisis moyennant certaines conditions supplémentaires, par exemple $\| \Psi_j \| = 1$. Le système d'éléments orthogonaux vérifiant cette condition est dit *orthonormé*.

Nous avons déjà rencontré le système de polynômes

$$T_n(x) = 2^{n-1} x^n + \dots,$$

orthogonaux sur l'intervalle fermé $[-1, 1]$ avec le poids $2/(\pi\sqrt{1-x^2})$ et noté à cette occasion que leurs valeurs sont calculables au moyen de la formule de récurrence

$$T_{n+1}(x) = 2xT_n(x) - T_{n-1}(x). \tag{8}$$

La recherche des polynômes orthogonaux de Tchebycheff selon (8) est à préférer au calcul direct suivant la formule explicite (7). En effet,

1. Dans le premier cas il est inutile de conserver en mémoire ou de calculer les coefficients b_{jl}.

2. D'ordinaire, on doit trouver les valeurs de tous les polynômes $\psi_1(x), \dots, \psi_n(x)$ à la fois et en un même point. Avec la formule (7) on calcule chaque polynôme séparément, ce qui fait que le calcul comprend en tout $\asymp n^2$ opérations arithmétiques.

$a(n) \asymp b(n)$ signifie que les deux éléments sont d'un même ordre de grandeur, i.e.

$$a(n) = O(b(n)), \quad b(n) = O(a(n)).$$

Tandis que dans le cas de la formule (8) ces opérations sont au nombre de $O(n)$.

3. Les valeurs de $T_n(x)$ obtenues directement par la formule (7) risquent d'être affectées d'une grande erreur de calcul: $T_n(x)$ est la somme

$$T_n(x) = \sum_{j=0}^{n} d_{nj} x^j. \tag{9}$$

En calcul automatique, les termes $d_{nj}x^j$ peuvent acquérir une erreur absolue de l'ordre de $|d_{nj}x^j| 2^{-t}$, d'où une erreur en $D_n(x) 2^{-t}$ sur $T_n(x)$ avec

$$D_n(x) = \sum_{j=0}^{n} |d_{nj}x^j|.$$

Minorons $D_n(x)$; l'égalité

$$T_n(|x|i) = \sum_{j=0}^{n} d_{nj}(|x|i)^j$$

entraîne l'estimation

$$|T_n|(|x|i) \leqslant \sum_{j=0}^{n} |d_{nj}x^j| = D_n(x).$$

D'autre part, selon (8.4) on a, pour x réel,

$$T_n(|x|i) = \frac{(|x|+\sqrt{1+|x|^2})^n + (|x|-\sqrt{1+|x|^2})^n}{2} i^n;$$

vu que

$$(|x|+\sqrt{1+|x|^2})(|x|-\sqrt{1+|x|^2}) = -1,$$

on en tire

$$D_n(x) \geqslant |T_n(|x|i)| \geqslant \frac{(|x|+\sqrt{1+|x|^2})^n - (|x|+\sqrt{1+|x|^2})^{-n}}{2}.$$

Ainsi, si n sont grands et $x \neq 0$, le calcul direct à l'aide de la formule (7) introduit une erreur de l'ordre de

$$(|x| + \sqrt{1+|x|^2})^n 2^{-t}.$$

PROBLÈME. Démontrer l'égalité

$$D_n(x) = |T_n(|x|i)|.$$

Une fois de plus nous sommes en présence de perte de chiffres significatifs: $|T_n(x)| \leqslant 1$, mais en calculant $T_n(x)$ avec (9) on

l'obtient comme somme de termes de signe variable grands en valeur absolue, qui présente donc une erreur importante.

On peut montrer par ailleurs que le calcul de $T_n(x)$ par la formule de récurrence (8) fournit une erreur de l'ordre de $O(n^2 2^{-t})$. On comprend désormais toute la portée des relations de récurrence du type (8) reliant les valeurs des polynômes orthogonaux relativement aux autres fonctions de poids $p(x)$.

Passons à des formations plus générales. Soit L un opérateur auto-adjoint sur l'espace hilbertien H: $\varphi_1 \in H$, $\varphi_j = L^{j-1}\varphi_1$; supposons que les éléments $\varphi_1, \ldots, \varphi_n$ sont linéairement indépendants. Selon le lemme 1 de ce paragraphe, on peut construire un système d'éléments orthogonaux

$$\psi_j = \sum_{i=1}^{j} b_{ji}\varphi_j,$$

avec tous les b_{ij} égaux à l'unité.

THÉORÈME. *Les éléments ψ_j vérifient le système de relations*

$$\psi_{j+1} + \alpha_{j+1,\,j}\psi_j + \alpha_{j+1,\,j-1}\psi_{j-1} - L\psi_j = 0, \qquad (10)$$

$$\alpha_{j+1,\,j-1} > 0.$$

DÉMONSTRATION. On a

$$L\psi_j = \sum_{i=1}^{j} b_{ji} L\varphi_i = \sum_{i=1}^{j} b_{ji}\varphi_{i+1}.$$

Vu (6): $\varphi_{(j+1)} = A_{j+1}\psi_{(j+1)}$, on écrit

$$L\psi_j = \sum_{i=1}^{j+1} \alpha_{j+1,\,i}\psi_i. \qquad (11)$$

Notons que $\alpha_{j+1,\,j+1} = 1$ conformément à (3), (6).

En multipliant scalairement les deux membres de (11) par ψ_k, $k \leqslant j+1$, il vient

$$(\psi_j, L\psi_k) = (L\psi_j, \psi_k) = \sum_{i=1}^{j+1} \alpha_{j+1,\,i}(\psi_i, \psi_k) = \alpha_{j+1,\,k}(\psi_k, \psi_k). \qquad (12)$$

Exprimons $L\psi_k$ à l'aide de (11), puis portons cette expression dans le premier membre de la dernière égalité

$$\Big(\psi_j, \sum_{i=1}^{k+1} \alpha_{k+1,\,i}\psi_i\Big) = \alpha_{j+1,\,k}(\psi_k, \psi_k).$$

Pour $k+1 < j$, le premier membre s'annule, donc $\alpha_{j+1,\,k} = 0$ quand $k < j-1$. Et la relation (11) se met sous la forme

$$\alpha_{j+1,\,j+1}\psi_{j+1} + \alpha_{j+1,\,j}\psi_j + \alpha_{j+1,\,j-1}\psi_{j-1} - L\psi_j = 0, \qquad (13)$$

où, selon (12),

$$\alpha_{j+1,\,k} = (\psi_j,\ L\psi_k)/(\psi_k,\ \psi_k).$$

Nous avons

$$(\psi_j,\ L\psi_{j+1})/(\psi_{j+1},\ \psi_{j+1}) = \alpha_{j+1,\,j+1} = 1,$$

i.e. $(L\psi_j,\ \psi_{j+1}) = (\psi_{j+1},\ \psi_{j+1})$; donc

$$\alpha_{j+1,\,j-1} = (\psi_j,\ L\psi_{j-1})/(\psi_{j-1},\ \psi_{j-1}) = (\psi_j,\ \psi_j)/(\psi_{j-1},\ \psi_{j-1}) > 0.$$

D'autre part,

$$\alpha_{j+1,\,j} = (L\psi_j,\ \psi_j)/(\psi_j,\ \psi_j),$$

et la relation de récurrence (10) s'écrit

$$\psi_{j+1} + \frac{(L\psi_j,\ \psi_j)}{(\psi_j,\ \psi_j)}\psi_j + \frac{(\psi_j,\ \psi_j)}{(\psi_{j-1},\ \psi_{j-1})}\psi_{j-1} - L\psi_j = 0. \tag{14}$$

Le théorème est démontré.

Considérons maintenant un cas intéressant pour nous: le cas A) pour $\varphi_1 \equiv 1$ et L un opérateur de multiplication par x. ψ_j est alors le polynôme $P_{j-1}(x)$ de degré $j-1$ et, en changeant les notations $\alpha_{jk} = d_{j-1,\,k-1}$, $j = n+1$, s'écrit

$$P_{n+1}(x) + d_{n+1,\,n}P_n(x) + d_{n+1,\,n-1}P_{n-1}(x) - xP_n(x) = 0, \tag{15}$$

$$d_{n+1,\,n} = \frac{\int_a^b xp(x)P_n^2(x)\,dx}{\int_a^b p(x)P_n^2(x)\,dx}, \quad d_{n+1,\,n-1} = \frac{\int_a^b p(x)P_n^2(x)\,dx}{\int_a^b p(x)P_{n-1}^2(x)\,dx} > 0. \tag{16}$$

LEMME 2. *Soit* $\{\psi_j\}_n$ *un système d'éléments*

$$\psi_j = \sum_{i=1}^{j} b_{ji}\varphi_i \quad \textit{pour} \quad b_{jj} = 1,$$

construit en orthogonalisant successivement le système $\{\varphi_j\}_n$, *et soit* $\{z_j\}_n$ *un système orthogonal d'éléments*

$$z_j = \sum_{i=1}^{j} d_{ji}\varphi_i, \ \textit{tous les}\ d_{jj} = 1. \tag{17}$$

Alors $z_j = \psi_j$ *quels que soient* $j \leqslant n$.

DÉMONSTRATION. La proposition est vraie pour $j = 1$; supposons qu'elle l'est pour tous les $j \leqslant k$. (17) entraîne

$$z_{k+1} = \varphi_{k+1} - \sum_{i=1}^{k} \widetilde{a}_{k+1,i}z_i.$$

En multipliant par z_l, il vient

$$0 = (\varphi_{k+1},\ z_l) - \widetilde{a}_{k+1,\,l}(z_l,\ z_l);$$

d'où

$$\widetilde{a}_{k+1,\,i} = \frac{(\varphi_{k+1},\, \psi_i)}{(\psi_i,\, \psi_i)}\,.$$

En se rappelant la démonstration du lemme 1, on constate que z_{k+1} est également déterminé par l'égalité (5). Donc, $z_{k+1} = \psi_{k+1}$, ce qui démontre l'unicité du système orthogonal du type considéré.

§ 16. Polynômes orthogonaux

Nous allons énumérer les systèmes les plus répandus de polynômes orthogonaux en particularisant la fonction de poids.

1. *Polynômes de Jacobi.* L'intervalle est $[-1,1]$ et la fonction de poids

$$p(x) = (1-x)^{\alpha}(1+x)^{\beta}, \quad \alpha,\ \beta > -1.$$

Les polynômes

$$P_n^{(\alpha,\,\beta)}(x) = \frac{(-1)^n}{2^n n!}(1-x)^{-\alpha}(1+x)^{-\beta}\frac{d^n}{dx^n}\left((1-x)^{\alpha+n}(1+x)^{\beta+n}\right)$$

sont dits *polynômes de Jacobi*. On a les relations

$$\| P_n^{(\alpha,\,\beta)}(x) \| = \left(\frac{2^{\alpha+\beta+1}\Gamma(\alpha+n+1)\,\Gamma(\beta+n+1)}{n!\,(\alpha+\beta+2n+1)\,\Gamma(\alpha+\beta+n+1)}\right)^{1/2},$$

$$\begin{aligned}(\alpha+\beta+2n)(\alpha+\beta+2n+1)(\alpha+\beta+2n+2)\,x P_n^{(\alpha,\,\beta)}(x) = \\ = 2(n+1)(\alpha+\beta+n+1)(\alpha+\beta+2n)\,P_{n+1}^{(\alpha,\,\beta)}(x) + \\ + (\beta^2-\alpha^2)(\alpha+\beta+2n+1)\,P_n^{(\alpha,\,\beta)}(x) + \\ + 2(\alpha+n)(\beta+n)(\alpha+\beta+2n+2)\,P_{n-1}^{(\alpha,\,\beta)}(x), \qquad (1)\end{aligned}$$

où Γ est la fonction gamma d'Euler.

Les polynômes de Jacobi satisfont à l'équation différentielle

$$\begin{aligned}L_{\alpha,\,\beta}(P_n^{(\alpha,\,\beta)}(x)) \equiv (1-x^2)(P_n^{(\alpha,\,\beta)}(x))_{xx} + ((\beta-\alpha) - \\ -(\alpha+\beta+2)\,x)(P_n^{(\alpha,\,\beta)}(x))_x = -n(\alpha+\beta+n+1)\,P_n^{(\alpha,\,\beta)}(x),\end{aligned}$$

i.e. ce sont les fonctions propres de l'opérateur différentiel $L_{\alpha\beta}$.

2. *Polynômes de Legendre.* Les polynômes de Jacobi pour $\alpha = \beta = 0$, c'est-à-dire lorsque la fonction de poids $p(x) \equiv 1$, sont dits *polynômes de Legendre* et se notent

$$L_n(x) = \frac{1}{2^n n!}\frac{d^n}{dx^n}(x^2-1)^n$$

avec pour norme

$$\| L_n \| = \sqrt{2/(2n+1)}$$

et la relation de récurrence

$$(n+1)\,L_{n+1}(x) - (2n+1)\,xL_n(x) + nL_{n-1}(x) = 0.$$

On remarquera que pour n grands, cette relation, de même que (1) avec α et β quelconques, s'apparente à la relation de récurrence pour les polynômes de Tchebycheff.

3. *Polynômes de Tchebycheff.* Pour $\alpha = \beta = -1/2$, $p(x) = 1/\sqrt{1-x^2}$ les polynômes de Jacobi deviennent les *polynômes de Tchebycheff de première espèce* $T_n(x)$.

4. *Polynômes de Tchebycheff de deuxième espèce.* Pour $\alpha = \beta = 1/2$, $p(x) = \sqrt{1-x^2}$, les polynômes de Jacobi s'appellent *polynômes de Tchebycheff de deuxième espèce*

$$U_n(x) = (\sin((n+1)\arccos x))/\sqrt{1-x^2} = T'_{n+1}(x)/(n+1), \quad n = 0, 1, \ldots,$$

de norme $\| U_n \| = \pi/2$. La relation de récurrence est la même pour les polynômes de Tchebycheff des deux espèces.

5. *Polynômes d'Hermite.* L'intervalle est $(-\infty, \infty)$ et la fonction de poids $p(x) = \exp(-x^2)$. Ce sont les polynômes

$$H_n(x) = (-1)^n \exp(x^2) \frac{d^n}{dx^n} (\exp(-x^2))$$

avec pour norme

$$\| H_n \| = \sqrt{2^n \cdot n! \sqrt{\pi}}$$

et relation de récurrence

$$H_{n+1}(x) - 2xH_n(x) + 2nH_{n-1}(x) = 0.$$

Ils satisfont à l'équation différentielle

$$H''_n - 2xH'_n = -2nH_n.$$

6. *Polynômes de Laguerre.* L'intervalle est $[0, \infty)$ et la fonction de poids $p(x) = x^\alpha \exp(-x)$. Ce sont les polynômes

$$L_n^{(\alpha)}(x) = (-1)^n x^{-\alpha} \exp x \frac{d^n}{dx^n} (x^{\alpha+n} \exp(-x))$$

de norme

$$\| L_n^{(\alpha)} \| = \sqrt{n!\,\Gamma(\alpha+n+1)}$$

et de relation de récurrence

$$L_{n+1}^{(\alpha)}(x) - (x - \alpha - 2n - 1) L_n^{(\alpha)}(x) + n(\alpha+n) L_{n-1}^{(\alpha)}(x) = 0.$$

Les polynômes de Laguerre vérifient l'équation différentielle

$$x(L_n^{(\alpha)}(x))_{xx} + (\alpha + 1 - x)(L_n^{(\alpha)}(x))_x = -nL_n^{(\alpha)}(x).$$

Etablissons certaines propriétés des polynômes orthogonaux. Soit $P_0(x), \ldots, P_n(x), \ldots$ un système de polynômes orthogonaux sur le segment $[a, b]$:

$$P_n(x) = x^n + \sum_{j=0}^{n-1} p_{nj} x^j.$$

LEMME 1. *Tout polynôme $P_n(x)$ admet exactement n zéros distincts sur l'intervalle (a, b).*

DÉMONSTRATION. Supposons que $P_n(x)$ ne possède sur (a, b) que $l < n$ zéros $x_1, \ldots, x_l$ de multiplicité impaire. Alors

$$P_n(x) \prod_{j=1}^{l} (x - x_j),$$

ne change pas de signe sur $[a, b]$; aussi

$$\int_a^b p(x) P_n(x) \prod_{j=1}^{l} (x - x_j)\, dx \neq 0.$$

D'autre part, cette intégrale est égale à zéro, car $P_n(x)$ est orthogonal par rapport à tous les polynômes de degré inférieur. La contradiction obtenue démontre le lemme.

LEMME 2. *Soient $x_1^{(n)} < \ldots < x_n^{(n)}$ les zéros de $P_n(x)$. Alors les zéros de $P_{n-1}(x)$ et de $P_n(x)$ alternent, c'est-à-dire:*

$$a < x_1^{(n)} < x_1^{(n-1)} < \ldots < x_{n-1}^{(n-1)} < x_n^{(n)} < b. \tag{2}$$

DÉMONSTRATION. Supposons cette proposition vraie pour un certain n. Portons $x = x_i^{(n)}$ dans (15.15); il vient

$$P_{n+1}(x_i^{(n)}) + d_{n+1,\,n-1} P_{n-1}(x_i^{(n)}) = 0.$$

Puisque $d_{n+1,\,n-1} > 0$,

$$\operatorname{sg} P_{n+1}(x_i^{(n)}) = -\operatorname{sg} P_{n-1}(x_i^{(n)}). \tag{3}$$

Comme tous les zéros du polynôme $P_{n-1}(x) = x^{n-1} + \ldots$ appartiennent à l'intervalle (a, b), on a

$$\operatorname{sg} P_{n-1}(b) = 1, \quad \operatorname{sg} P_{n-1}(a) = (-1)^{n-1}.$$

On en tire, compte tenu de (2),

$$\operatorname{sg} P_{n-1}(x_i^{(n)}) = (-1)^{n-i}.$$

De même

$$\operatorname{sg} P_{n+1}(b) = 1, \quad \operatorname{sg} P_{n+1}(a) = (-1)^{n+1}.$$

Il résulte de la dernière relation et de (3) que le polynôme $P_{n+1}(x)$ change de signe en tout point de la suite $a, x_1^{(n)}, \ldots, x_n^{(n)}, b$ et qu'il admet un zéro entre deux points consécutifs quelconques, ce qui démontre la validité de (2) pour n suivant. On peut raisonner à partir de $n = 1$ parce que cette valeur vérifie toutes les relations écrites, y compris (3). Le lemme 2 est donc démontré.

Considérons un cas d'un intérêt pratique certain où le poids est une fonction paire par rapport au milieu du segment $[a, b]$. Posons, pour fixer les idées, $[a, b] = [-1, 1]$ et, partant, $p(x) = p(-x)$. Nous avons alors le

LEMME 3. *Tous les polynômes $P_{2l}(x)$ sont pairs et tous les $P_{2l-1}(x)$ impairs, c'est-à-dire*

$$P_j(-x) = (-1)^j P_j(x) \text{ quels que soient } j, \tag{4}$$

et la relation de récurrence (15.15) *s'écrit*

$$P_{j+1}(x) - xP_j(x) + d_{j+1,j-1}P_{j-1}(x) = 0. \tag{5}$$

DÉMONSTRATION. $p(x)$ étant pair, $P_0(x) = 1$ et $P_1(x) = x$ sont orthogonaux, et

$$(P_0, P_1) = \int_{-1}^{1} xp(x)\,dx = 0.$$

Ils sont donc éléments d'un système orthogonal. Supposons que (4) est démontrée pour $j \leqslant n$ et (5) pour $j < n$. Revenant à la formule (15.15), on a

$$P_{n+1}(x) + (d_{n+1,n} - x)P_n(x) + d_{n+1,n-1}P_{n-1}(x) = 0.$$

Puisque (4) est vraie et $p(x)$ pair, $xp(x)P_n^2(x)$ est impair, donc, conformément à (15.16), $d_{n+1,n} = 0$. Il en découle la validité de (5) quand $j = n$. La fonction $xP_n(x)$ est de même parité que $P_{n-1}(x)$; aussi $P_{n+1}(x)$, qui figure à gauche dans (5), pour $j = n$ est de même parité que $P_{n-1}(x)$. Nous avons démontré (4) pour la valeur suivante $j = n + 1$ et, avec elle, le lemme 3.

Il en résulte en particulier que dans le cas considéré les zéros des polynômes orthogonaux $P_n(x)$ sont symétriques par rapport à l'origine des coordonnées. En particulier pour n impair l'un de ces zéros est $x = 0$ du moment que $P_n(x)$ sont impairs.

Les relations de récurrence citées aux nos 2-4 pour des polynômes orthogonaux concrets ne sont pas identiques à (5), puisque cette dernière est vraie pour des polynômes orthogonaux normés de façon que le coefficient du terme de plus haut degré est égal à 1.

Signalons une propriété remarquable de la répartition des zéros de polynômes orthogonaux. Soit $[a, b] = [-1, 1]$ et supposons que le poids $p(x)$ est presque partout positif sur ce segment. Désignons par $w_n(x_1, x_2)$ le nombre de zéros de $P_n(x)$ appartenant au segment

$[x_1, x_2]$. On a

$$\lim_{n\to\infty} \frac{w_n(x_1, x_2)}{n} = \int_{x_1}^{x_2} \frac{2}{\pi\sqrt{1-x^2}}\, dx. \tag{6}$$

Ainsi, pour un poids quelconque, les zéros des polynômes orthogonaux sont asymptotiquement équirépartis avec la densité $\frac{2}{\pi\sqrt{1-x^2}}$.

§ 17. Dérivation numérique

Les formules les plus simples de *dérivation numérique* sont fournies par la dérivation des formules d'interpolation.

Etant données les valeurs d'une fonction aux points $x_1, \ldots, x_n$, on demande de calculer la dérivée $f^{(k)}(x)$. Formons le polynôme d'interpolation $L_n(x)$ et posons $f^{(k)}(x) \approx L_n^{(k)}(x)$. On effectue les mêmes substitutions à l'aide des polynômes de Bessel, d'Everett, etc.

Evaluons l'erreur dans les formules de dérivation numérique. On a

$$f(x) - L_n(x) = f(x; x_1; \ldots; x_n)\, \omega_n(x);$$

d'où, d'après la formule de Leibniz,

$$f^{(k)}(x) - L_n^{(k)}(x) = \sum_{j=0}^{k} C_k^j (f(x; x_1; \ldots; x_n))^{(j)}\, \omega_n^{(k-j)}(x). \tag{1}$$

Soit $g(x)$ q fois continûment dérivable. En vertu de (5.4) qui donne la dérivée moyennant la différence divisée, on a

$$g(x; x+\varepsilon; \ldots; x+q\varepsilon) = g^{(q)}(x_\varepsilon)/q!,$$

avec $x \leqslant x_\varepsilon \leqslant x + q\varepsilon$; d'où

$$g^{(q)}(x) = q! \lim_{\varepsilon\to 0} g(x; x+\varepsilon; \ldots; x+q\varepsilon).$$

Par conséquent,

$$(f(x; x_1; \ldots; x_m))^{(q)} = q! \left[\lim_{\varepsilon\to 0} f(x; x+\varepsilon; \ldots; x+ \right.$$
$$\left. +q\varepsilon; x_1; \ldots; x_m)\right]. \tag{2}$$

En vertu de la définition de la différence divisée à points multiples, l'expression entre crochets est

$$f(\underbrace{x; \ldots; x}_{q+1 \text{ fois}}; x_1; \ldots; x_m).$$

Donc, (1) se récrit

$$f^{(k)}(x)-L_n^{(k)}(x)=\sum_{j=0}^{k}\frac{k!}{(k-j)!}f(\underbrace{x;\ \ldots;\ x}_{j+1\text{ fois}};\ x_1;\ \ldots;\ x_n)\,\omega_n^{(k-j)}(x). \tag{3}$$

Nous avons exprimé l'erreur sur les formules de dérivation numérique au moyen des différences divisées de la fonction. Après avoir évalué celles-ci par les dérivées conformément à (5.4), on aboutit à l'estimation d'erreur

$$|f^{(k)}(x)-L_n^{(k)}(x)|\leqslant$$
$$\leqslant\sum_{j=0}^{k}\frac{k!}{(k-j)!\,(n+j)!}\max_{[y_1,\,y_2]}|f^{(n+j)}(x)|\,|\omega_n^{(k-j)}(x)|, \tag{4}$$

où, comme plus haut,

$$y_1=\min(x,\,x_1,\,\ldots,\,x_n),\quad y_2=\max(x,\,x_1,\,\ldots,\,x_n).$$

Considérons quelques cas particuliers. Supposons que la dérivée k-ième soit déterminée par la dérivation d'un polynôme d'interpolation de degré k aux $k+1$ abscisses $x_1,\ \ldots,\ x_{k+1}$. Du moment que

$$L_{k+1}(x)=f(x_1)+\ldots+f(x_1;\ \ldots;\ x_{k+1})\,(x-x_1)\ \ldots\ (x-x_k),$$

la relation $f^{(k)}(x)\approx L_{k+1}^{k}(x)$ s'écrit

$$f^{(k)}(x)\approx k!f(x_1;\ \ldots;\ x_{k+1}),$$

i.e. elle est une conséquence de (5.4). Pour $k=1$, (3) donne

$$f'(x)-L_n'(x)=f(x;\ x_1;\ \ldots;\ x_n)\,\omega_n'(x)+$$
$$+f(x;\ x;\ x_1;\ \ldots;\ x_n)\,\omega_n(x)=\frac{f^{(n)}(\zeta_1)}{n!}\,\omega_n'(x)+$$
$$+\frac{f^{(n+1)}(\zeta_2)}{(n+1)!}\,\omega_n(x),\quad \zeta_1,\ \zeta_2\in[y_1,\,y_2]. \tag{5}$$

Si x est l'un des points $x_1,\ \ldots,\ x_n$ utilisés dans l'interpolation, alors $\omega_n(x)=0$, et la relation (5) entraîne

$$|f'(x)-L_n'(x)|\leqslant\Big(\max_{[y_1,\,y_2]}|f^{(n)}(x)|\Big)\,\frac{|\omega_n'(x)|}{n!}.$$

Considérons le cas des points équidistants : $x_j-x_{j-1}\equiv h$. $\omega_n^{(j)}(x)$ est la somme du produit de $n-j$ facteurs de l'ordre de h chacun, si bien que l'ordre de $\omega_n^{(j)}(x)$ est $O(h^{n-j})$. Le second membre de (1) est donc

$$f(x;\ x_1;\ \ldots;\ x_n)\,\omega_n^{(k)}(x)+O(h^{n-k+1})=O(h^{n-k}).$$

Si $\omega_n^{(k)}(x)=0$ en x, il ressort de cette relation que l'ordre de grandeur de l'erreur augmente. Citons quelques formules usuelles possédant cette propriété.

Proposons-nous de calculer $f''(x)$ en x_0 pour les abscisses $x_{-l}, \ldots, x_0, \ldots, x_l$, $x_j - x_{j-1} \equiv h$. Alors $n = 2l + 1$, $\omega_n(x) = (x - x_{-l}) \ldots (x - x_l)$ est impaire par rapport à x_0. Aussi $\omega_n(x_0) = \omega_n''(x_0) = 0$, et la relation (3) s'écrit pour $k = 2$:

$$f''(x_0) - L_n''(x_0) = 2f(x; x; x_{-l}; \ldots; x_l)\, \omega_n'(x_0).$$

La dérivation directe fournit

$$\omega_n'(x_0) = (-1)^l (l!)^2 h^{2l}.$$

Ainsi, compte tenu de (5.4), il vient

$$f''(x_0) - L_n''(x_0) = \frac{2(-1)^l (l!)^2}{(2l+2)!} f^{(2l+2)}(\zeta) h^{2l}. \qquad [(6)$$

$L_n''(x_0)$ s'exprime par les différences f_0^{2j}. On a finalement

$$f''(x_0) \approx L_n''(x_0) = h^{-2} \sum_{j=1}^{l} \frac{2(-1)^{j-1} ((j-1)!)^2}{(2j)!} f_0^{2j} \qquad (7)$$

avec (6) pour reste.

Les cas les plus usités de (7) sont $l = 1$, $l = 2$:

$$f''(x_0) \approx f_0^2/h^2, \quad f''(x_0) \approx \left(f_0^2 - \frac{1}{12} f_0^4\right)\Big/ h^2. \qquad (8)$$

Le calcul de toute dérivée $f^{(2k)}(x_0)$ sur l'ensemble ci-dessus de points d'interpolation s'effectue avec la même élévation de 1 de l'ordre de l'erreur par rapport à h^{n-2k}.

Trouvons la dérivée $y'(x)$ en $\bar{x} = x_0 + h/2$ pour les points

$$x_{-(l-1)}, \ldots, x_0, x_1, \ldots, x_l, x_{j+1} - x_j \equiv h.$$

Dans ce cas la fonction $\omega_n(x)$ est paire par rapport à $\bar{x}$, et la relation (5) s'écrit

$$f'(\bar{x}) - L_n'(\bar{x}) = \frac{f^{(n+1)}(\zeta)}{(n+1)!} \omega_n(\bar{x}).$$

Comme dans le cas de la dérivée seconde, écrivons $L_n'(\bar{x})$ en fonction de $f_{1/2}^{2j+1}$ et calculons $\omega_n(\bar{x})$. On aboutit à la formule de dérivation numérique

$$f'(x_0 + h/2) \approx L_{2l}'(x_0 + h/2) =$$

$$= h^{-1} \sum_{j=0}^{l-1} \frac{(-1)^j ((1/2)(3/2) \ldots (j-1/2))^2}{(2j+1)!} f_{1/2}^{2j+1} \qquad (9)$$

dont le reste est

$$\frac{(-1)^l ((1/2)(3/2) \ldots (l-1/2))^2}{(2l+1)!} f^{(2l+1)}(\zeta) h^{2l}.$$

Ecrivons les formules les plus simples (correspondant à $l=1$, $l=2$)

$$f'(x_0+h/2) \approx f^1_{1/2}/h,$$
$$f'(x_0+h/2) \approx \left(f^1_{1/2} - \frac{1}{24} f^3_{1/2}\right)\Big/h. \tag{10}$$

Pour notre ensemble de points, l'ordre de précision pour une dérivée quelconque d'ordre impair s'élèvera de 1.

Pour conclure, parlons des *formules unilatérales de dérivation numérique* qui offrent un intérêt pratique. Soit à calculer $f'(x_0)$ par interpolation aux abscisses $x_0, \ldots, x_n$ à pas constant. Selon (5), on a

$$f'(x_0) - L'_{n+1}(x_0) = \frac{f^{(n+1)}(\zeta)}{(n+1)!}\,\omega'_n(x_0). \tag{11}$$

En y portant $L_{n+1}(x)$ sous forme de polynôme de Newton pour l'interpolation en avant, il vient

$$f'(x_0) \approx L'_{n+1}(x_0) = h^{-1} \sum_{j=1}^{n} \frac{(-1)^{j-1}}{j}\, f^j_{j/2} \tag{12}$$

avec pour reste

$$\frac{(-1)^n}{(n+1)} f^{(n+1)}(\zeta)\, h^n.$$

Les formules les plus simples (correspondant à $l=1$, $l=2$) sont

$$f'(x_0) \approx f^1_{1/2}/h, \tag{13}$$
$$f'(x_0) \approx \left(f^1_{1/2} - \frac{1}{2} f^2_1\right)\Big/h.$$

On construit par analogie les formules si les points d'interpolation sont $x_0, x_{-1}, \ldots, x_{-n}$:

$$f'(x_0) \approx L'_{n+1}(x_0) = h^{-1} \sum_{j=1}^{n} \frac{1}{j}\, f^j_{-j/2}, \tag{14}$$

le reste étant

$$\frac{1}{n+1} f^{(n+1)}(\zeta)\, h^n.$$

§ 18. Erreur de calcul dans les formules de dérivation numérique

Il peut arriver, dans une dérivation numérique, que la diminution de l'erreur de méthode fasse du même coup augmenter l'influence des erreurs provenant des données initiales et du calcul même.

Supposons, pour fixer les idées, que $f'(x_0)$ se définit par la relation

$$f'(x_0) \approx (f(x_1) - f(x_0))/h. \tag{1}$$

Conformément à (17.11) son reste est

$$r_1 = -f''(\zeta)\, h/2.$$

Supposons que $|f''(\zeta)| \leqslant M_2$ et, donc, $|r_1| \leqslant M_2 h/2$.

Si le calcul de $f(x_i)$ conduit à une certaine erreur ε_i ($|\varepsilon_i| \leqslant E$), celui de $f'(x_0)$ donnera un terme d'erreur supplémentaire

$$r_2 = (\varepsilon_1 - \varepsilon_0)/h, \quad |r_2| \leqslant 2E/h.$$

Le souci de la simplicité nous fait négliger les arrondis lors du calcul du second membre de (1). On a l'évaluation d'erreur

$$|r| \leqslant |r_1| + |r_2| \leqslant g(h) = M_2 h/2 + 2E/h. \tag{2}$$

L'erreur est petite si h l'est, or la diminution de h entraîne la croissance du second terme (fig. 2.18.1). L'équation $g'(h) = 0$ donne le point d'extrémum de $g(h)$:

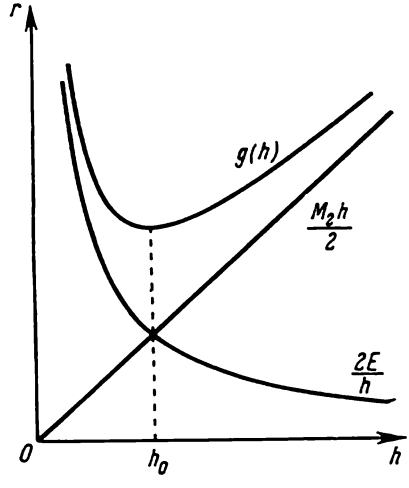


Fig. 2.18.1

$$h_0 = 2\sqrt{E/M_2},$$

puis

$$g(h_0) = 2\sqrt{M_2 E}.$$

Ainsi, pour aucun h on ne peut affirmer que l'erreur sur le résultat sera de l'ordre de $o(\sqrt{E})$.

ε_i sont dues aux erreurs affectant les valeurs données des fonctions, si par exemple ces dernières sont définies par des mesures ou calculées par une formule approchée. Puisque ces valeurs sont arrondies une deuxième fois à l'entrée, on estime que $E \geqslant \text{const}\ 2^{-t}$, t étant la longueur du nombre traité. Donc, dans le meilleur des cas la moitié des chiffres de $f'(x_0)$ est exacte.

Le recours aux formules plus exactes améliore quelque peu cet état de chose. Calculons la dérivée $y^{(k)}(x)$ moyennant la formule

$$y^{(k)}(x) \approx \Big(\sum_0^n c_j y_j\Big)/h^k, \quad c_j = O(1), \tag{3}$$

le reste étant $O(h^l)$. Dans le cas $k = 1, 2$, les formules de calcul des dérivées d'ordre k contiennent, on l'a vu, h^k au dénominateur.

L'erreur générée par les erreurs du second membre de (3) est évaluée par const $\cdot E/h^k$. On a donc au lieu de (2)

$$|r| \leqslant g_1(h) = A_1 h^l + A_2 E/h^k.$$

Le second membre atteint un minimum pour un h de l'ordre de $E^{1/(l+k)}$ et son ordre de grandeur est $E^{l/(l+k)}$. On voit que l'ordre de l'erreur augmente devant E avec l. Ceci étant, la valeur du pas qui correspond au minimum de l'évaluation d'erreur augmente constamment. N'oublions surtout pas que A_1 et A_2 peuvent croître avec l, donc il n'y a lieu d'augmenter l que jusqu'à une certaine valeur.

Dans la pratique on recourt volontiers à toutes sortes de méthodes de lissage des valeurs initiales des fonctions, qui permettent, d'une part, de diminuer E et, d'autre part, de rendre la fonction plus lisse et, partant, d'utiliser les formules de dérivation par interpolation plus exactes. Nous en reparlerons au chapitre V.

Bibliographie du chapitre II

§ 1 : I [15], II [22] ; §§ 4-7, 10 : II [7, 36] ; § 18; III [25, 34].

CHAPITRE III

INTÉGRATION NUMÉRIQUE

§ 1. Formules de quadrature de Newton-Cotes

Les formules de quadrature les plus simples s'obtiennent par des considérations intuitives. Soit à calculer l'intégrale

$$I = \int_a^b f(x)\,dx.$$

Si $f(x) \approx$ const sur le segment considéré, on pose $I \approx (b-a) \times \times f(\zeta)$, ζ étant un point arbitraire de $[a, b]$. On prend tout naturellement pour ζ le point médian du segment; on obtient alors la *formule des rectangles*

$$I \approx (b-a) f\left(\frac{b+a}{2}\right).$$

Supposons que $f(x)$ soit proche d'une fonction linéaire, ce qui permet de remplacer l'intégrale par l'aire d'une trapèze de hauteur $(b-a)$ et de bases $f(a)$ et $f(b)$ (fig. 3.1.1). On aboutit à la *formule des trapèzes*

$$I \approx (b-a)\frac{f(a)+f(b)}{2}.$$

Avec la supposition ci-dessus sur $f(x)$, des considérations géométriques permettent d'espérer des résultats satisfaisants avec la formule des rectangles. En effet, $(b-a) f\left(\frac{b+a}{2}\right)$ représente l'aire d'une trapèze quelconque de hauteur $b-a$ et de médiane $f\left(\frac{b+a}{2}\right)$, en particulier, celle d'une trapèze dont l'un des côtés se confond avec la tangente à la courbe $y = f(x)$ au point $\left(\frac{b+a}{2}, f\left(\frac{b+a}{2}\right)\right)$ (fig. 3.1.2).

Les formules de quadrature plus compliquées se construisent, comme celles de dérivation numérique, à l'aide de l'appareil d'interpolation. Cette même technique nous permettra également d'évaluer l'erreur commise sur les formules de quadrature.

On demande de calculer l'intégrale

$$\int_a^b f(x)\,p(x)\,dx.$$

Le lecteur verra plus bas (§ 3) les raisons qui nous incitent à introduire la fonction $p(x)$ dite *fonction de poids*. Fixons-nous certains

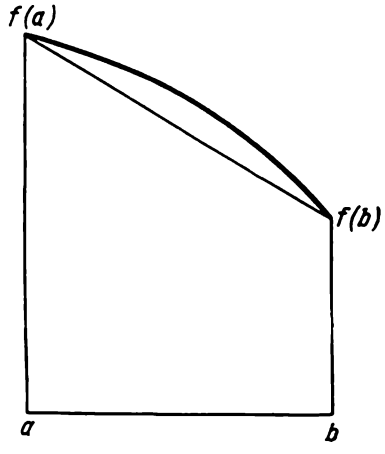


Fig. 3.1.1

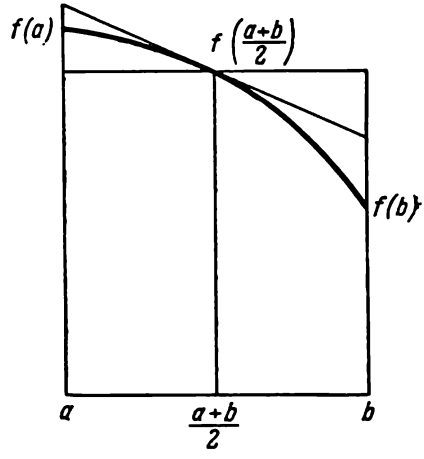


Fig. 3.1.2

$d_1, \ldots, d_n \in [-1, 1]$ et formons un polynôme d'interpolation $L_n(x)$ de degré $n-1$ qui coïncide avec $f(x)$ aux points

$$x_j = \frac{b+a}{2} + \frac{b-a}{2} d_j.$$

Posons

$$\int_a^b f(x)\, p(x)\, dx \approx \int_a^b L_n(x)\, p(x)\, dx.$$

On a

$$R_n(f) = \int_a^b f(x)\, p(x)\, dx - \int_a^b L_n(x)\, p(x)\, dx = \int_a^b p(x)\,(f(x) - L_n(x))\, dx.$$

On évalue la différence $f(x) - L_n(x)$ en recourant à l'estimation de l'erreur sur le polynôme de Lagrange

$$|f(x) - L_n(x)| \leqslant (\max_{[a,\,b]} |f^{(n)}(x)|)\, \frac{|\omega_n(x)|}{n!},$$

avec

$$\omega_n(x) = (x - x_1) \ldots (x - x_n).$$

D'où

$$|R_n(f)| \leqslant (\max_{[a,\,b]} |f^{(n)}(x)|) \int_a^b \frac{|\omega_n(x)|\, |p(x)|}{n!}\, dx.$$

Effectuons le changement de variable dans la dernière intégrale

$$x = X(t) = \frac{b+a}{2} + \frac{b-a}{2}\, t;$$

on trouve

$$\frac{1}{n!}\int_a^b |\omega_n(x)\, p(x)|\, dx = D(d_1, \ldots, d_n)\left(\frac{b-a}{2}\right)^{n+1},$$

où

$$D(d_1, \ldots, d_n) = \int_{-1}^{1} \frac{|\omega_n^0(t)\, p^0(t)|}{n!}\, dt,$$

$$\omega_n^0(t) = (t-d_1)\ldots(t-d_n),\quad p^0(t) = p\left(\frac{b+a}{2} + \frac{b-a}{2}\, t\right).$$

On a donc l'estimation

$$|R_n(f)| \leqslant D(d_1, \ldots, d_n)\left(\max_{[a,\,b]} |f^{(n)}(x)|\right)\left(\frac{b-a}{2}\right)^{n+1}. \tag{1}$$

Supposons tous les d_j distincts. Alors

$$L_n(x) = \sum_{j=1}^{n} f(x_j) \prod_{i \neq j} \frac{x-x_i}{x_j-x_i}.$$

En faisant $x = X(t)$ il vient

$$\int_a^b p(x)\, L_n(x)\, dx = \left(\frac{b-a}{2}\right) \sum_{j=1}^{n} D_j f(x_j), \tag{2}$$

avec

$$D_j = \int_{-1}^{1} p^0(t) \prod_{i \neq j} \frac{t-d_i}{d_j-d_i}\, dt.$$

Ainsi, on aboutit à une formule de quadrature de la forme

$$\int_a^b f(x)\, p(x)\, dx \approx \frac{b-a}{2} \sum_{j=1}^{n} D_j f\left(\frac{b+a}{2} + \frac{b-a}{2}\, d_j\right). \tag{3}$$

De même que dans la dérivation numérique, pour un problème jouissant d'une certaine symétrie, la méthode possédant une symétrie du même type présente souvent des avantages supplémentaires.

LEMME. *Soient $p(x)$ une fonction de poids paire par rapport au milieu du segment $[a, b]$ et x_j des points symétriques par rapport à ce point médian: $d_j = -d_{n+1-j}$. Alors les coefficients de la formule de quadrature associés aux points symétriques sont égaux:*

$$D_j = D_{n+1-j}. \tag{4}$$

DÉMONSTRATION. Ecrivons les coefficients comme suit :

$$D_j = \int_{-1}^{1} p^0(t) \frac{\omega(t)}{\omega'(d_j)(t-d_j)}\, dt,$$

où $\omega(t) = (t-d_1) \ldots (t-d_n)$. Pour n pair on a $\omega(t) = \omega(-t)$ et $\omega'(d_j) = -\omega'(d_{n+1-j})$. Pour n impair, il vient $\omega(t) = -\omega(-t)$ et $\omega'(d_j) = \omega'(d_{n+1-j})$. Dans les deux cas la substitution $t = -t'$ donne

$$D_j = \int_{-1}^{1} p^0(t') \frac{\omega(t')}{\omega'(d_{n+1-j})(t'+d_j)}\, dt' =$$

$$= \int_{-1}^{1} p^0(t') \frac{\omega(t')}{\omega'(d_{n+1-j})(t'-d_{n+1-j})}\, dt' = D_{n+1-j}. \qquad (5)$$

Ces formules de quadrature « symétriques » possèdent une propriété complémentaire que leur construction formelle ne laissait pas entrevoir, à savoir elles sont exactes pour toute fonction impaire par rapport au milieu du segment $[a, b]$. En effet, $p(x)$ étant pair, ces fonctions vérifient $\int_a^b p(x) f(x)\, dx = 0$, et, conformément à (4), on a $\sum_{j=1}^{n} D_j f(x_j) = 0$; aussi $R_n(f) = 0$. En particulier ces formules sont exactes pour tout monôme $\text{const} \cdot \left(x - \frac{b+a}{2}\right)^{2l+1}$.

Considérons une formule pareille pour n impair. Elle est stricte pour $f(x) = \text{const} \cdot \left(x - \frac{b+a}{2}\right)^n$ et, par construction, pour un polynôme quelconque de degré $n-1$. Elle est donc stricte pour n'importe quel polynôme de degré n. Par conséquent, les formules de quadrature ainsi construites basées sur $2q-1$ et $2q$ points et telles que les points sont symétriques par rapport au milieu du segment s'avèrent exactes pour des polynômes de degré $2q-1$.

Afin de préciser l'erreur dans les formules de quadrature à points en nombre impair en l'évaluant par $f^{(n+1)}(x)$ et non par $f^{(n)}(x)$, on remplace la fonction à intégrer par le polynôme d'interpolation de Lagrange avec $(b+a)/2$ pour point double. On établit ci-dessous, pour $p(x) \equiv 1$ des formules de quadrature élémentaires avec évaluation de l'erreur commise; si $n = 1$ et $n = 3$ dans les formules « symétriques », l'évaluation se fait à l'aide de $f^{(n+1)}(x)$.

1. *Formule des rectangles.* Soit $n = 1$, $d_1 = 0$. On a alors

$$D = \int_{-1}^{1} |t|\, dt = 1, \quad D_1 = \int_{-1}^{1} 1 \cdot dt = 2$$

et la formule de quadrature

$$\int_a^b f(x)\,dx \approx (b-a) f\left(\frac{a+b}{2}\right) \tag{6}$$

avec pour estimation du reste

$$\max_{[a,b]} |f'(x)| \frac{(b-a)^2}{4}.$$

2. *Formule des rectangles à point multiple.* Soit $n=2$, $d_1 = d_2 = 0$. On a

$$D = \int_{-1}^{1} \frac{t^2}{2}\,dt = \frac{1}{3},$$

$$L_2(x) = f\left(\frac{a+b}{2}\right) + f'\left(\frac{a+b}{2}\right)\left(x - \frac{a+b}{2}\right),$$

$$\int_a^b L_2(x)\,dx = (b-a) f\left(\frac{a+b}{2}\right).$$

Ainsi, la formule de quadrature s'écrit

$$\int_a^b f(x)\,dx \approx (b-a) f\left(\frac{a+b}{2}\right).$$

le reste étant évalué par

$$\max_{[a,b]} |f''(x)| \frac{(b-a)^3}{24}.$$

La formule est la même dans les deux cas, mais l'estimation du reste diffère.

3. *Formule des trapèzes.* Soit $n=2$, $d_1 = -1$, $d_2 = 1$. On a

$$D = \int_{-1}^{1} \frac{|t^2-1|}{2}\,dt = \frac{2}{3},$$

$$D_1 = \int_{-1}^{1} \frac{1-t}{2}\,dt = 1,$$

$$D_2 = \int_{-1}^{1} \frac{1+t}{2}\,dt = 1.$$

La *formule des trapèzes* se met sous la forme

$$\int_a^b f(x)\,dx \approx \frac{(b-a)}{2}(f(a)+f(b)) \tag{7}$$

et son reste est évalué par

$$\max_{[a, b]} |f''(x)| \frac{(b-a)^3}{12}.$$

4. *Formule de Simpson.* Soit $n=4$, $d_1=-1$, $d_2=d_4=0$, $d_3=1$. Alors

$$D=\int_{-1}^{1} \frac{t^2(1-t^2)}{4!}\,dt=\frac{1}{90}.$$

Conformément à la formule d'interpolation à points multiples on écrit

$$L_4(x)=L_3(x)+f\left(a;\ b;\ \frac{a+b}{2};\ \frac{a+b}{2}\right)(x-a)(x-b)\times$$
$$\times\left(x-\frac{a+b}{2}\right),$$

où

$$L_3(x)=f(a)+f(a;\ b)(x-a)+f\left(a;\ b;\ \frac{a+b}{2}\right)(x-a)(x-b).$$

Le second terme de l'expression de $L_4(x)$ est une fonction impaire par rapport au milieu du segment $[a, b]$. C'est pourquoi

$$\int_a^b L_4(x)\,dx=\int_a^b L_3(x)\,dx.$$

$L_3(x)$ est un polynôme d'interpolation de degré 2 correspondant à

$$d_1=-1,\quad d_2=0,\quad d_3=1$$

(fig. 3.1.3). Il correspond à ces d_1, d_2, d_3 :

$$D_1=\frac{1}{3},\quad D_2=\frac{4}{3},\quad D_3=\frac{1}{3}.$$

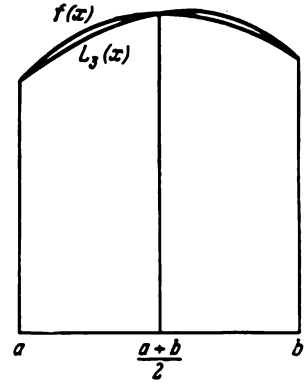


Fig. 3.1.3

La formule de quadrature

$$\int_a^b f(x)\,dx\approx\frac{b-a}{6}\left(f(a)+4f\left(\frac{b+a}{2}\right)+f(b)\right) \tag{8}$$

est dite *formule de Simpson.* On évalue son reste par

$$\max_{[a, b]} |f^{(4)}(x)| \frac{(b-a)^5}{2880}.$$

5. Une formule de quadrature exacte pour des polynômes du troisième degré s'obtient par exemple pour $n=4$ si l'on pose

$$d_1=-1, \quad d_2=-\frac{1}{3}, \quad d_3=\frac{1}{3}, \quad d_4=1.$$

Le schéma décrit nous conduit alors à la formule

$$\int_a^b f(x)\,dx \approx \frac{b-a}{8}\left(f(a)+3f\left(a+\frac{b-a}{3}\right)+\right.$$
$$\left.+3f\left(a+\frac{2(b-a)}{3}\right)+f(b)\right. \qquad (9)$$

Pour le même degré des polynômes la vérifiant, elle exige un point de plus que celle de Simpson, on a donc moins de raisons de s'en servir.

Une remarque s'impose à propos de la formule (9). Puisque $f(x)-L_n(x)=f^{(n)}(\zeta(x))\frac{\omega_n(x)}{n!}$, le reste se met sous la forme

$$\int_a^b f^{(n)}(\zeta(x))\frac{\omega_n(x)}{n!}\,dx.$$

Dans les cas examinés plus haut $2-4\omega_n(x)$ ne présentait pas de changement de signe sur $[a, b]$. Selon le théorème de la moyenne

$$\int_a^b h(x)\,g(x)\,dx=h(\bar{x})\int_a^b g(x)\,dx, \quad \bar{x}\in[a, b], \text{ si } g(x)\geqslant 0;$$

ce qui peut s'écrire

$$f^{(n)}(\tilde{x})\int_a^b \frac{\omega_n(x)}{n!}\,dx. \qquad (10)$$

Portant dans (10) une expression concrète de $\omega_n(x)$ on représente l'erreur dans les formules des rectangles, des trapèzes et de Simpson respectivement par

$$R_2(f)=f''(\tilde{x}_1)\frac{(b-a)^3}{24},$$

$$R_3(f)=-f''(\tilde{x}_2)\frac{(b-a)^3}{12}, \quad R_4(f)=-f^{(4)}(\tilde{x}_3)\frac{(b-a)^5}{2880}.$$

Si $f^{(n)}(x)$ varie peu sur $[a, b]$, l'estimation (1) ne sera pas en l'occurrence trop forte. Pour (9)

$$\omega_n(x)=(x-a)\left(x-\frac{2a+b}{3}\right)\left(x-\frac{a+2b}{3}\right)(x-b)$$

change de signe sur $[a, b]$ et le reste ne se met pas directement sous la forme (10).

Il se peut donc que, dans le cas de cette formule de quadrature, l'estimation (1) soit trop forte par comparaison avec le résultat exact. Nous procéderons au § 2 à des constructions générales qui permettront une évaluation exacte de l'erreur correspondante au moyen du maximum de $|f^{(4)}(x)|$.

Le présent chapitre se propose essentiellement de traiter des procédés de calcul d'intégrales de fonctions données analytiquement et de mettre au point des principes présidant à l'élaboration de programmes standard d'intégration de fonctions de ce type. La théorie des formules de quadrature a naturellement affaire à d'autres problèmes liés par exemple au traitement des données expérimentales.

Arrêtons-nous à titre d'exemple sur les *formules de Tchebycheff* auxquelles on recourait volontiers pour calculer la jauge des navires. La position du problème qui a abouti à ces formules rappelle beaucoup la planification d'expériences (ch. II, § 1). On demande de calculer $\int_{-1}^{1} f(x)\,dx$, la fonction $f(x)$ pouvant être approchée avec une précision admissible par un polynôme de degré q. L'obtention de chaque $f(x)$ par mesure est un procédé assez coûteux et est entachée d'une erreur accidentelle relativement importante. Supposons que les erreurs de mesure sont indépendantes, ont la même variance et une espérance mathématique nulle. La variance de la valeur approchée $S_n(f)$ calculée par la formule de quadrature

$$I(f) \approx S_n(f) = \sum_{j=1}^{n} c_j f(x_j)$$

est alors

$$\text{Var} \sum_{j=1}^{n} c_j^2.$$

La condition $I(f) = S_n(f)$ pour $f = \text{const}$ s'écrit

$$2 = \sum_{j=1}^{n} c_j. \tag{11}$$

On vérifie sans peine que le minimum de $\text{Var} \sum_{j=1}^{n} c_j^2$ dans la condition (11) est réalisé pour $c_1 = \ldots = c_n = 2/n$. Ces raisonnements nous conduisent au problème suivant: parmi toutes les formules de quadrature

$$I(f) \approx \frac{2}{n} \sum_{j=1}^{n} f(x_j),$$

exactes pour des polynômes de degré q, trouver celle qui correspond au plus petit n possible. Pour $q = 0$ et $q = 1$, c'est la formule des rectangles $I(f) = 2f(0)$; pour $q = 2$, $q = 3$, c'est la formule de Gauss (voir § 3)

$$I(f) \approx f\left(-\frac{1}{\sqrt{3}}\right) + f\left(\frac{1}{\sqrt{3}}\right).$$

§ 2. Evaluation de l'erreur dans une formule de quadrature sur une classe de fonctions

Considérons la formule de quadrature

$$I(f)=\int_a^b f(x)\,dx \approx S_N(f)=(b-a)\sum_{j=1}^{N} D_j f(x_j)\ ^*),$$

dont on sait qu'elle est exacte pour tous les polynômes de degré $n-1$. Soit $C_n(A_n, [a, b])$ une classe de fonctions $n-1$ fois continûment dérivables de dérivée $f^{(n)}(x)$ continue par morceaux telle que $|f^{(n)}(x)| \leqslant A_n$ si $x \in [a, b]$. On se propose d'expliciter

$$\sup_{f \in C_n(A_n[a,b])} |I(f)-S_N(f)|,$$

c'est-à-dire évaluer l'erreur avec un maximum de précision pour la classe considérée.

Signalons pour mémoire la formule de Taylor avec reste sous forme intégrale

$$f(x)=T_c^n f+\int_c^x f^{(n)}(y)\,\frac{(x-y)^{n-1}}{(n-1)!}\,dy, \tag{1}$$

avec

$$T_c^n f=\sum_{j=0}^{n-1} f^{(j)}(c)\,\frac{(x-c)^j}{j!}.$$

Portons cette expression de f dans celle de l'erreur

$$R_N(f)=I(f)-S_N(f). \tag{2}$$

$R_n(f)$ peut se mettre sous la forme

$$R_N(f)=R_N(T_c^n f)+R_N\left(\int_c^x f^{(n)}(y)\,\frac{(x-y)^{n-1}}{(n-1)!}\,dy\right).$$

Le premier terme à droite est nul en vertu de l'hypothèse que la formule de quadrature est exacte pour les polynômes de degré $n-1$. Ainsi,

$$R_N(f)=\int_a^b\left(\int_c^x f^{(n)}(y)\,\frac{(x-y)^{n-1}}{(n-1)!}\,dy\right)dx-$$

$$-(b-a)\sum_{j=1}^{N} D_j\int_c^{x_j} f^{(n)}(y)\,\frac{(x_j-y)^{n-1}}{(n-1)!}\,dy.$$

*) Dans la formule de quadrature le signe somme est précédé soit du facteur $b-a$, soit de $(b-a)/2$; le sens des coefficients D_j varie en conséquence.

Soit $c = a$ pour fixer les idées. En permutant les variables d'intégration x et y de la première intégrale, on trouve

$$\int_a^b \int_a^x f^{(n)}(y) \frac{(x-y)^{n-1}}{(n-1)!}\, dy\, dx =$$

$$= \int_a^b f^{(n)}(y) \left(\int_y^b \frac{(x-y)^{n-1}}{(n-1)!}\, dx \right) dy = \int_a^b f^{(n)}(y) \frac{(b-y)^n}{n!}\, dy.$$

Chaque $\int_a^{x_j} f^{(n)}(y) \frac{(x_j-y)^{n-1}}{(n-1)!}\, dy$ se prête à l'écriture suivante

$$\int_a^b f^{(n)}(y) \frac{\overline{(x_j-y)}^{n-1}}{(n-1)!}\, dy,$$

$$\bar{t} = \begin{cases} t & \text{pour } t > 0, \\ 0 & \text{pour } t \leqslant 0. \end{cases}$$

Donc,

$$R_N(f) = \int_a^b K_N(y) f^{(n)}(y)\, dy, \tag{3}$$

où

$$K_N(y) = \frac{(b-y)^n}{n!} - (b-a) \sum_{j=1}^{N} D_j \frac{\overline{(x_j-y)}^{n-1}}{(n-1)!}. \tag{4}$$

Il en résulte l'évaluation de l'erreur

$$|R_N(f)| \leqslant A_n \int_a^b |K_N(y)|\, dy. \tag{5}$$

Montrons qu'elle ne peut être améliorée. Sur tout segment $[a, x_1]$, $[x_1, x_2]$, . . ., $[x_N, b]$ la fonction $K_N(y)$ est un polynôme de degré n et admet donc n zéros au plus. Aussi sur le segment $[a, b]$ tout entier $K_N(y)$ en possède $n(N+1)$ au plus. Faisons $z(x) = A_n \operatorname{sg} K_N(x)$. La fonction

$$Z(x) = \int_a^x z(y) \frac{(x-y)^{n-1}}{(n-1)!}\, dy$$

admet pour la dérivée d'ordre n la fonction continue par morceaux $z(x)$, $|z(x)| = A_n$, c'est-à-dire qu'elle appartient à la classe con-

sidérée. On a

$$R_N(Z) = \int_a^b K_N(y)\, A_n \operatorname{sg} K_N(y)\, dy = A_n \int_a^b |K_N(y)|\, dy,$$

par conséquent,

$$\sup_{f \in C_n(A_n, [a,b])} |R_N(f)| = A_n \int_a^b |K_N(y)|\, dy. \tag{6}$$

PROBLÈME. Examinons une classe de fonctions n fois continûment dérivables, $|f^{(n)}(x)| \leqslant A_n$. Démontrer que sur cette classe on a également

$$\sup |R_N(f)| = A_n \int_a^b |K_N(y)|\, dy. \tag{7}$$

Evaluons l'erreur commise dans la formule de quadrature

$$\int_a^b f(x)\, dx \approx \frac{b-a}{8}\left(f(a) + 3f\left(\frac{2a+b}{3}\right) + 3f\left(\frac{a+2b}{3}\right) + f(b)\right)$$

par $\max_{[a,b]} |f^{(4)}(x)|$. Effectuons la substitution $y = a + (b-a)\,t$ dans l'intégrale $\int_a^b |K_N(y)|\, dy$. Il vient

$$\int_a^b |K_N(y)|\, dy = (b-a)^{n+1} \int_0^1 |K_N^0(t)|\, dt,$$

où

$$K_N^0(t) = \frac{(1-t)^n}{n!} - \sum_{j=1}^{N} D_j \frac{\overline{(d_j^0 - t)}^{n-1}}{(n-1)!}, \quad d_j^0 = \frac{x_j - a}{b-a}. \tag{8}$$

Dans notre cas

$$K_4^0(t) = \frac{(1-t)^4}{24} - \frac{1}{48}\left(\overline{(0-t)}^3 + 3\left(\overline{\frac{1}{3} - t}\right)^3 + 3\left(\overline{\frac{2}{3} - t}\right)^3 + (1-t)^3\right);$$

notons que $\overline{(0-t)}^3 \equiv 0$ si $0 \leqslant t$. On vérifie que pour $0 \leqslant t \leqslant 1$ on a $K_4^0(t) \leqslant 0$. L'évaluation d'erreur exige ensuite qu'on calcule

$$\int_0^1 |K_4^0(t)|\, dt = -\int_0^1 K_4^0(t)\, dt.$$

Si $K_N^0(t)$ conserve son signe, l'erreur est évaluée par une technique plus commode.

LEMME. *Admettons que* $K_N^0(t)$ *conserve son signe sur* $[a, b]$. *Alors*

$$R_N(f) = f^{(n)}(\zeta)\frac{(b-a)^{n+1}}{n!}\left(\int_0^1 t^n\,dt - \sum_{j=1}^{N} D_j\,(d_j^0)^n\right). \tag{9}$$

DÉMONSTRATION. En appliquant la formule de Lagrange généralisée au second membre de (3), on obtient

$$R_N(f) = f^{(n)}(\zeta)\int_a^b K_N(y)\,dy = f^{(n)}(\zeta)\,(b-a)^{n+1}\int_0^1 K_N^0(t)\,dt.$$

Portons-y $f(y) = f_0(y) = \dfrac{(y-a)^n}{n!}$, nous avons

$$R_N(f_0) = (b-a)^{n+1}\int_0^1 K_N^0(t)\,dt.$$

Portant la même fonction dans l'égalité

$$R_N(f_0) = I(f_0) - S_N(f_0),$$

on est conduit à

$$R_N(f_0) = \frac{(b-a)^{n+1}}{n!}\left(\int_0^1 t^n\,dt - \sum_{j=1}^{N} D_j\,(d_j^0)^n\right).$$

Ces relations entraînent

$$\int_0^1 K_N^0(t)\,dt = \frac{1}{n!}\left(\int_0^1 t^n\,dt - \sum_{j=1}^{N} D_j\,(d_j^0)^n\right) = \sigma.$$

La validité de (9) est immédiate. Notons que σ est l'erreur dans la formule de quadrature $\displaystyle\int_0^1 f(x)\,dx \approx \sum_{j=1}^{N} D_j f(d_j^0)$ lorsqu'on calcule $\displaystyle\int_0^1 \frac{x^n}{n!}\,dx$.

Ainsi, dans le cas examiné

$$R_N(f) = f^{(4)}(\zeta)\frac{(b-a)^5}{24}\left(\frac{1}{5} - \frac{3}{8}\left(\frac{1}{3}\right)^4 - \frac{3}{8}\left(\frac{2}{3}\right)^4 - \frac{1}{8}\right) =$$

$$= -f^{(4)}(\zeta)\frac{(b-a)^5}{6480}.$$

Pour cinq points on construit, comme on l'a fait pour la formule de Simpson, la formule suivante

$$\int_a^b f(x)\,dx \approx \frac{b-a}{90}\left(12 f\left(\frac{a+b}{2}\right) + 32\left(f\left(\frac{3a+b}{4}\right) + f\left(\frac{a+3b}{4}\right)\right) + 7\,(f(a)+f(b))\right) \tag{10}$$

avec pour reste $-\dfrac{1}{1\,935\,360}(b-a)^7 f^{(6)}(\zeta)$.

Pour déduire ce reste on suppose le point $\frac{a+b}{2}$ multiple, la fonction $K_5^0(y)$ conserve le signe et on évalue l'erreur suivant le schéma décrit.

On recourt rarement aux formules possédant un grand nombre de points équidistants.

§ 3. Formules de quadrature de Gauss

La formule de Simpson et celle des rectangles attestent qu'un choix heureux des points élève le degré des polynômes vérifiant exactement la formule de quadrature.

C'est pourquoi on pose le problème d'optimisation suivant. Pour un nombre donné n de points, construire la formule de quadrature

$$I(f)=\int_a^b f(x)\,p(x)\,dx \approx S_n(f)=\frac{b-a}{2}\sum_{j=1}^{n} D_j f(x_j), \tag{1}$$

exacte pour les polynômes de plus haut degré possible. Il s'agit des *formules de Gauss*.

Cette façon de poser la question s'explique par des considérations suivantes. Supposons que nous avons construit la formule (1) stricte pour tout polynôme P_m de degré m:

$$I(P_m) = S_n(P_m). \tag{2}$$

Ceci étant, on a pour un P_m quelconque la suite d'égalités

$$R_n(f) = R_n(P_m) + R_n(f - P_m) = R_n(f - P_m).$$

On a visiblement

$$|R_n(g)| \leqslant \int_a^b |g(x)|\,|p(x)|\,dx + \frac{b-a}{2}\sum_{j=1}^{n} |D_j|\,|g(x_j)| \leqslant V_n \sup_{[a,b]} |g(x)|,$$

avec

$$V_n = \int_a^b |p(x)|\,dx + \left(\frac{b-a}{2}\right)\sum_{j=1}^{n} |D_j|. \tag{3}$$

Il en découle

$$|R_n(f)| \leqslant V_n \sup_{[a,b]} |f(x) - P_m(x)|.$$

Prenant à droite la borne inférieure sur tous les polynômes de degré m, on trouve l'estimation

$$|R_n(f)| \leqslant V_n E_m(f), \tag{4}$$

où

$$E_m(f) = \inf_{P_m} \sup_{[a, b]} |f(x) - P_m(x)|.$$

La fonction est mieux approchée par des polynômes de degré élevé, ce qui autorise à penser qu'une formule de quadrature exacte pour de pareils polynômes donne lieu à une erreur moins importante.

Ce qui est faux en général. Premièrement, l'évaluation (4) n'est pas précise, et, deuxièmement, le coefficient V_n est fonction de la formule même et, rigoureusement parlant, il peut croître avec le degré des polynômes pour lesquels la formule est exacte. Cependant, toute fausse qu'elle soit, notre supposition justifie dans une certaine mesure la façon dont nous avons posé le problème.

Si la fonction à intégrer de l'intégrale initiale $\int_a^b F(x)\,dx$ est mal approchée par des polynômes et que $F(x)/p(x)$ avec $p(x)$ une fonction élémentaire le soit bien, alors pour notre cas la formule de type (1) qui est exacte pour les polynômes d'un certain degré offre des avantages par rapport à la formule

$$\int_a^b F(x)\,dx \approx \frac{b-a}{2} \sum_{j=1}^{n} c_j F(x_j), \tag{5}$$

stricte pour les polynômes de même degré.

Si, lorsque $p(x) \neq 0$, on pose dans (1) $p(x) f(x) = F(x)$, il vient

$$\int_a^b F(x)\,dx \approx \frac{b-a}{2} \sum_{j=1}^{n} C_j F(x_j), \quad C_j = \frac{D_j}{p(x_j)}.$$

Cette formule de quadrature est évidemment exacte pour toutes les fonctions $p(x) P_m(x)$, $P_m(x)$ étant un polynôme de degré m. Ainsi, notre position initiale équivaut à rechercher une formule de type (5) qui soit exacte pour les fonctions $F(x) = p(x) P_m(x)$ avec un m maximal.

Revenons à la première position du problème. Pour

$$P_m(x) = \sum_{q=0}^{m} a_q x^q,$$

on a

$$R_n(P_m) = \sum_{q=0}^{m} a_q R_n(x^q).$$

Ainsi, pour que la formule de quadrature (1) soit exacte pour les polynômes de degré m, il faut et il suffit qu'elle le soit pour toutes les fonctions x^q, $q = 0, \ldots, m$. Par conséquent, on doit avoir

les relations

$$R_n(x^q) = \int_a^b x^q p(x)\,dx - \frac{b-a}{2} \sum_{j=1}^{N} D_j x_j^q = 0, \quad q = 0, \ldots, m. \qquad (6)$$

Nous avons un système de $m + 1$ équations par rapport aux inconnues $x_1, \ldots, x_n, D_1, \ldots, D_n$.

Si $m \leqslant 2n - 1$, le nombre d'équations est au plus égal à celui d'inconnues, et on s'attend que le système algébrique (6) possède une solution. En résolvant ce système on peut essayer de construire des formules de quadrature correspondant à $m = 2n - 1$, sans pouvoir affirmer que leurs points appartiennent au segment $[a, b]$. Si cette dernière condition n'est pas satisfaite, la fonction $f(x)$ risque de ne pas être déterminée aux points de base de l'intégration et l'opération de quadrature s'avère impossible.

Signalons qu'au chapitre VIII, en élaborant des méthodes des différences finies de résolution des équations différentielles ordinaires on sera mis en présence de formules de quadrature basées sur des points extérieurs au segment $[a, b]$.

Formons les formules de quadrature correspondant à la valeur maximale $m = 2n - 1$.

LEMME 1. *Si $x_1, \ldots, x_n$ sont les points de base de la formule de quadrature* (1) *exacte pour tous les polynômes de degré $2n - 1$, alors*

$$\int_a^b \omega_n(x) P_{n-1}(x) p(x)\,dx = 0,$$

où $\omega_n(x) = (x - x_1) \ldots (x - x_n)$ et $P_{n-1}(x)$ est un polynôme quelconque de degré $n - 1$ au plus.

DÉMONSTRATION. Soit $P_{n-1}(x)$ un polynôme de degré inférieur ou égal à $n - 1$. Par hypothèse, la formule (1) vérifie

$$Q_{2n-1}(x) = \omega_n(x) P_{n-1}(x).$$

Aussi

$$\int_a^b \omega_n(x) P_{n-1}(x) p(x)\,dx = \int_a^b Q_{2n-1}(x) p(x)\,dx =$$

$$= \frac{b-a}{2} \sum_{j=1}^{n} D_j Q_{2n-1}(x_j) = 0;$$

la dernière relation résulte de $Q_{2n-1}(x_j) = 0$ quels que soient j. Le lemme 1 est démontré.

On suppose ensuite que $p(x) > 0$ presque partout sur $[a, b]$.

Les raisonnements du § 15, ch. II entraînent l'unicité du poly-

nôme $\psi_n(x) = x^n + \ldots$ orthogonal par rapport à tous les polynômes de degré inférieur pour le produit scalaire défini par

$$(f, g) = \int_a^b f(x)\,\overline{g}(x)\,p(x)\,dx.$$

C'est pourquoi $\psi_n(x) = \omega_n(x)$ et les points de base de la formule de quadrature cherchée doivent coïncider avec les zéros de $\psi_n(x)$; selon § 15, ch. II le polynôme $\psi_n(x)$ admet sur $[a, b]$ n zéros distincts.

LEMME. 2. *Soient $x_1, \ldots, x_n$ les zéros de $\psi_n(x)$ et* (1) *une formule de quadrature exacte pour les polynômes de degré $n-1$. Alors* (1) *l'est également pour les polynômes de degré $2n-1$.*

DÉMONSTRATION. Mettons un polynôme $Q_{2n-1}(x)$ arbitraire de degré $2n-1$ sous la forme

$$Q_{2n-1}(x) = \psi_n(x)\,g_{n-1}(x) + r_{n-1}(x),$$

où g_{n-1} et r_{n-1} sont des polynômes de degré $n-1$. On a

$$R_n(Q_{2n-1}) = R_n(\psi_n g_{n-1}) + R_n(r_{n-1}) = R_n(\psi_n g_{n-1}),$$

$$R_n(\psi_n g_{n-1}) = \int_a^b \psi_n(x)\,g_{n-1}(x)\,p(x)\,dx -$$

$$-\frac{b-a}{2}\sum_{j=1}^{n} D_j\psi_n(x_j)\,g_{n-1}(x_j) = 0,$$

du moment que

$$\int_a^b \psi_n(x)\,g_{n-1}(x)\,p(x)\,dx = 0,$$

$\psi_n(x)$ étant orthogonal par rapport aux polynômes de degré inférieur et tous les $\psi_n(x_j) = 0$ par hypothèse; alors $R_n(Q_{2n-1}) = 0$. Le lemme 2 est démontré.

Il est maintenant possible de construire la formule nécessaire. Donnons-nous les points d'interpolation $x_1, \ldots, x_n$, $\psi_n(x_j) = 0$, et établissons la formule de Newton-Cotes suivant le schéma du § 1. Nous obtenons

$$\int_a^b f(x)\,p(x)\,dx \approx \frac{b-a}{2}\sum_{j=1}^{n} G_j f(x_j), \tag{7}$$

qui est stricte pour les polynômes de degré $2n-1$.

Si $p(x)$ est presque partout supérieure à zéro, il n'existe pas de formule de quadrature stricte pour tous les polynômes de degré $2n$. En effet, soit $Q_{2n}(x) = (x-x_1)^2 \ldots (x-x_n)^2$; alors le premier

membre de (1) devient

$$\int_a^b ((x-x_1)\dots(x-x_n))^2\, p(x)\,dx > 0,$$

et le deuxième s'annule.

Etudions certaines propriétés de (7) et démontrons en particulier le signe plus des coefficients G_j. La fonction $\left(\frac{\psi_n(x)}{x-x_l}\right)^2$ est un polynôme de degré $2n-2$ s'annulant en tous les points $x_j \neq x_l$. La formule (7) sera vérifiée par cette fonction, donc,

$$\int_a^b \left(\frac{\psi_n(x)}{x-x_l}\right)^2 p(x)\,dx = \frac{b-a}{2} G_l \left(\frac{\psi_n(x)}{x-x_l}\right)^2\Big|_{x_l}.$$

En développant $\frac{\psi_n(x)}{x-x_l}$ il vient

$$G_l = \frac{2}{b-a}\int_a^b \left(\prod_{j\neq l}\frac{x-x_j}{x_l-x_j}\right)^2 p(x)\,dx > 0.$$

La formule (7) est exacte pour $f(x)=1$, aussi

$$\int_a^b p(x)\,dx = \frac{b-a}{2}\sum_{j=1}^n G_j;$$

G_j étant positifs, la dernière égalité se récrit:

$$\int_a^b p(x)\,dx = \frac{b-a}{2}\sum_{j=1}^n |G_j|.$$

En revenant aux formules (3), (4) on trouve

$$V_n = \int_a^b p(x)\,dx + \frac{b-a}{2}\sum_{j=1}^n |G_j| = 2\int_a^b p(x)\,dx$$

(4) entraîne alors l'estimation

$$|R_n(f)| \leqslant 2\left(\int_a^b p(x)\,dx\right) E_{2n-1}(f). \tag{8}$$

Citons une autre estimation de l'erreur dans les formules de Gauss qui s'exprime à l'aide de

$$\max_{[a,\,b]} |f^{(2n)}(x)|.$$

Soit $L_{2n}(x)$ un polynôme d'interpolation de Lagrange basé sur les points doubles $x_1, \dots, x_n$, i.e.

$$L_{2n}(x_j) = f(x_j), \quad L'_{2n}(x_j) = f'(x_j).$$

Selon (2.6.11), on a

$$f(x)-L_{2n}(x)=f(x\,;x_1\,;x_1\,;\ldots\,;x_n\,;x_n)\,\psi_n^2(x).$$

Puisque pour les formules de Gauss $R_n(L_{2n}(x))=0$, on a

$$R_n(f)=R_n(f(x\,;x_1\,;x_1\,;\ldots\,;x_n\,;x_n)\,\psi_n^2(x))=$$

$$=\int_a^b f(x\,;x_1\,;x_1\,;\ldots\,;x_n\,;x_n)\,\psi_n^2(x)\,p(x)\,dx-$$

$$-\frac{b-a}{2}\sum_{j=1}^{n} D_j f(x_j\,;x_1\,;x_1\,;\ldots\,;x_n\,;x_n)\,\psi_n^2(x_j).$$

Le dernier terme est égal à zéro car $\psi_n(x_j)=0$ quels que soient j. C'est pourquoi

$$R_n(f)=\int_a^b f(x\,;x_1\,;x_1,\ldots,x_n,x_n)\,\psi_n^2(x)\,p(x)\,dx.$$

$\psi_n^2(x)\,p(x)$ étant non négative, on applique le théorème de la moyenne, ce qui donne

$$R_n(f)=f(\zeta\,;x_1\,;x_1\,;\ldots\,;x_n\,;x_n)\int_a^b \psi_n^2(x)\,p(x)\,dx\,;$$

exprimant $f(\zeta\,;x_1\,;x_1\,;\ldots\,;x_n\,;x_n)$ en fonction de la dérivée, il vient

$$R_n(f)=f^{(2n)}(\bar{\zeta})\int_a^b \frac{\psi_n^2(x)}{(2n)!}\,p(x)\,dx. \tag{9}$$

Si l'on veut se servir des formules de Gauss, il faut connaître leurs points et leurs coefficients. Pour $p(x)$ paire par rapport au point $\frac{a+b}{2}$, il résulte des considérations du § 16, ch. II que les zéros des polynômes orthogonaux, i.e. les points des formules de quadrature de Gauss, sont symétriques par rapport au milieu du segment $[a, b]$. En vertu du lemme du § 1, les coefficients de la formule d'interpolation de Gauss (7) vérifient la condition de parité de $G_j = G_{n+1-j}$. D'où une réduction de moitié des tableaux des formules de Gauss.

Si $p(x)\equiv 1$, alors G_j et $d_j=\frac{2x_j-(b+a)}{b-a}$ ne dépendent pas du segment $[a, b]$. En effet si le polynôme $\psi_n(x)=(x-x_1)\ldots(x-x_n)$ appartient à un système de polynômes orthogonaux de poids 1 sur $[a, b]$, alors le polynôme $(t-d_1)\ldots(t-d_n)$ est un élément d'un système de polynômes orthogonaux de poids 1 sur $[-1, 1]$. Aussi

ce polynôme, ses zéros et, conformément à (1.2), les coefficients D_j sont-ils déterminés de façon unique et indépendamment du segment $[a, b]$.

En dressant les tableaux des formules de quadrature de Gauss pour des poids différents, on utilisera avec profit la propriété d'alternance des zéros des polynômes orthogonaux $\psi_n(x)$ que nous avons démontrée au § 16, ch. II. Selon cette propriété, une fois qu'on a déterminé les racines de $\psi_{n-1}(x)$ on connaît les intervalles renfermant les racines de $\psi_n(x)$.

Signalons pour mémoire un jeu de formules de Gauss pour le segment $[-1, 1]$ et $p(x) = 1$. Le reste (9) est alors

$$f^{(2n)}(\zeta)\frac{2^{2n+1}(n!)^4}{(2n!)^3(2n+1)}.$$

La propriété de symétrie nous fait nous borner aux d_j non négatifs et à leurs coefficients:

Tableau 1

	d_1 D_1	d_2 D_2	d_3 D_3
1	0,0000000000 2,0000000000		
2	0,5773502692 1,0000000000		
3	0,0000000000 0,8888888888	0,7745966692 0,5555555556	
4	0,8611363115 0,3478548451	0,3399810436 0,6521451549	
5	0,0000000000 0,5688888888	0,9061798459 0,4786286705	0,5384693101 0,2369268851
6	0,9324695142 0,1713244924	0,6612093864 0,3607615730	0,2386191861 0,4679139346

On dispose actuellement, pour les points et les poids des formules de quadrature de Gauss pour $n \leqslant 512$, des tableaux à 20 décimales.

Voyons comment appliquer les constructions précédentes pour déduire d'autres formules de quadrature. Pour simplifier l'exposé faisons $[a, b] = [-1, 1]$.

Formules de Markov (ou *de Lobatto*). Supposons que le calcul de la formule

$$I(f) = \int_{-1}^{1} f(x)\,dx \approx \sum_{j=1}^{n} D_j f(d_j)$$

soit soumis à la condition supplémentaire: $d_1 = -1$, $d_n = 1$. Les paramètres libres dont dépend la formule sont alors $2n - 2$, et on a des raisons d'espérer résoudre le système (6) pour $m = 2n - 3$, c'est-à-dire construire une formule vérifiée par tous les polynômes de degré $2n - 3$. Le polynôme d'interpolation du premier degré coïncidant avec $f(x)$ en ± 1 est

$$Q_1(x) = f(-1)\frac{1-x}{2} + f(1)\frac{1+x}{2}.$$

Mettons l'intégrale de départ sous la forme

$$I(f) = \int_{-1}^{1} Q_1(x)\,dx + \int_{-1}^{1} \frac{f(x) - Q_1(x)}{1-x^2}\,p(x)\,dx, \tag{10}$$

avec $p(x) = 1 - x^2$.

Construisons la formule de Gauss basée sur $n - 2$ points relativement à la fonction de poids $p(x)$:

$$\int_{-1}^{1} f(x)\,p(x)\,dx \approx \sum_{j=2}^{n-1} \overline{D}_j f(d_j). \tag{11}$$

Pour la commodité nous avons désigné les points par $d_2, \ldots, d_{n-1}$. Posons

$$\int_{-1}^{1} f(x)\,dx \approx \int_{-1}^{1} Q_1(x)\,dx + \sum_{j=2}^{n-1} \overline{D}_j \frac{f(d_j) - Q_1(d_j)}{1-d_j^2}. \tag{12}$$

Après avoir substitué $Q_1(x)$ et calculé la première intégrale à droite, on obtient la formule de quadrature cherchée

$$\int_{-1}^{1} f(x)\,dx \approx D_1 f(-1) + \sum_{j=2}^{n-1} D_j f(d_j) + D_n f(1),$$

où, en particulier,

$$D_j = \overline{D}_j (1 - d_j^2)^{-1}, \qquad j = 2, \ldots, n-1.$$

En effet, soit $f(x)$ un polynôme de degré $2n - 3$. Le calcul de la première intégrale de (10) ne donne lieu à aucune erreur. La différence $f(x) - Q_1(x)$ est un polynôme de degré $2n - 3$ qui s'annule pour $x = \pm 1$; aussi $(f(x) - Q_1(x))(1 - x^2)^{-1}$ est un polynôme de degré $2n - 5$. Il vérifie la formule (11), et en substituant à la deuxième intégrale de (10) le deuxième terme à droite dans (12), on ne commet pas d'erreur. Pour $n = 2$, la formule de Markov se confond avec la formule des trapèzes, pour $n = 3$ avec celle de Simpson.

Soit maintenant $p(x)$ une fonction impaire sur $[-1, 1]$, $p(x) > 0$ presque partout sur $[0, 1]$, les points sont en nombre pair: $n = 2l$.

Si $p(x)$ est paire, on a $D_j = D_{n+1-j}$, $d_j = -d_{n+1-j}$ dans les formules précises.

Si $p(x)$ est impaire, il paraît naturel de rechercher des formules pareilles sous la forme

$$I(f) \approx \sum_{j=1}^{l} D_j (f(d_j) - f(-d_j)), \quad 0 < d_1 < \ldots < d_l \leqslant 1. \quad (13)$$

Une telle formule est nécessairement précise pour $f(x)$ paire. La formule (13) contient $n = 2l$ paramètres libres, ce qui fait espérer une formule de quadrature exacte pour les polynômes $x, \ldots, x^{2n-1}$. Puisque cette dernière formule est automatiquement vérifiée par les monômes $1, \ldots, x^{2n}$, elle l'est également par des polynômes de degré $2n$, c'est-à-dire supérieur de 1 à celui des polynômes dans le cas des formules de Gauss pour un poids positif.

Pour construire une formule de ce type, mettons l'intégrale initiale sous la forme

$$I(f) = \int_0^1 (f(x) - f(-x))\, p(x)\, dx = \int_0^1 \frac{f(x) - f(-x)}{x}\, xp(x)\, dx.$$

Le changement de variable $x = \sqrt{y}$ fournit

$$I(f) = \int_0^1 F(y)\, \bar{p}(y)\, dy$$

avec

$$F(y) = \frac{f(\sqrt{y}) - f(-\sqrt{y})}{2\sqrt{y}}, \quad \bar{p}(y) = p(\sqrt{y}).$$

Etablissons la formule de Gauss

$$\int_0^1 F(y)\, \bar{p}(y)\, dy \approx \sum_{j=1}^{l} \bar{D}_j F(\bar{d}_j), \quad (14)$$

exacte pour $F(y)$ de degré $2l-1$. Revenons à l'ancienne variable x; nous avons

$$I(f) \approx \sum_{j=1}^{l} \bar{D}_j \frac{f(\sqrt{\bar{d}_j}) - f(-\sqrt{\bar{d}_j})}{2\sqrt{\bar{d}_j}} = \sum_{j=1}^{l} D_j (f(d_j) - f(-d_j)).$$

Si $f(x) = \sum_{j=0}^{2n} a_j x^j$, alors

$$F(y) = \sum_{j=0}^{n-1} a_{2j+1} y^j.$$

Aussi (14) et la relation suivante deviennent-elles des égalités strictes.

Ces formules de quadrature permettent par exemple de calculer les intégrales $\int_0^\infty f(x)\cos x\,dx$; l'intégrale est décomposée en parties:

$$\int_{\pi n}^{(n+1)\pi} f(x)\cos x\,dx=(-1)^{n-1}\frac{\pi}{2}\int_{-1}^{1} f\left((x+2n+1)\frac{\pi}{2}\right)\sin\frac{\pi x}{2}\,dx,$$

et les dernières intégrales se calculent moyennant les formules (13) correspondant à $p(x)=\sin\frac{\pi x}{2}$.

Les *formules d'Hermite*

$$I(f)=\int_{-1}^{1}\frac{f(x)}{\sqrt{1-x^2}}\,dx\approx S_n(f)=\frac{\pi}{n}\sum_{j=1}^{n} f(x_j), \tag{15}$$

où $x_j=\cos\frac{(2j-1)\pi}{2n}$ sont les zéros du polynôme de Tchebycheff $T_n(x)$, sont d'un grand intérêt pratique, en particulier pour les problèmes d'approximation des fonctions. Il s'agit des formules de Gauss sur le segment $[-1, 1]$ avec pour poids $\frac{1}{\sqrt{1-x^2}}$. Vérifions que la formule de quadrature (15) est exacte pour les polynômes $T_m(x)$ si $m<2n$. On a pour $m=0$

$$I(T_0)=\int_{-1}^{1}\frac{1}{\sqrt{1-x^2}}\,dx=\pi,\quad S_n(T_0)=\pi$$

et $I(T_0)=S_n(T_0)$. Pour $m>0$, l'orthogonalité des fonctions $T_m(x)$ de poids $\frac{1}{\sqrt{1-x^2}}$ entraîne

$$I(T_m)=I(T_mT_0)=(T_m, T_0)=0.$$

D'autre part,

$$S_n(T_m)=\frac{\pi}{n}\sum_{j=1}^{n}\cos(m\arccos x_j)=\frac{\pi}{n}\sum_{j=1}^{n}\cos\frac{m(2j-1)\pi}{2n}=$$

$$=\frac{\pi}{2n}\sum_{j=1-n}^{n}\exp\left(\frac{m(2j-1)\pi i}{2n}\right).$$

Les quantités additionnées forment une progression géométrique de raison $\exp\left(\frac{m\pi i}{n}\right)$; pour $0<m<2n$ cette raison est différente de l'unité. Avec la formule exprimant la somme d'une progression

géométrique :

$$\sum_{j=1}^{n} aq^{j-1} = \frac{a(q^n-1)}{q-1},$$

on a

$$S_n(T_m) = \frac{\pi}{2n} \frac{\exp\left(\frac{m(2n+1)\pi i}{2n}\right) - \exp\left(\frac{m(1-2n)\pi i}{2n}\right)}{\exp\frac{m\pi i}{n} - 1}.$$

Puisque

$$\exp\left(\frac{m(2n+1)\pi i}{2n}\right) = \exp\left(\frac{m(1-2n)\pi i}{2n}\right),$$

$S_n(T_m) = 0 = I(T_m)$ pour $0 < m < 2n$. Du moment que la formule de quadrature est stricte pour tous les polynômes de Tchebycheff $T_m(x)$ avec $m < 2n$, elle l'est pour n'importe quel polynôme de degré $2n - 1$.

On dispose à l'heure actuelle de nombreux tableaux des formules de Gauss et des formules de type Markov, en particulier pour

$$[a, b] = [-1, 1], \quad p(x) = 1,$$
$$[a, b] = [0, 1] \qquad p(x) = x^{\alpha},$$
$$[a, b] = [0, 1] \qquad p(x) = (x(1-x))^{\alpha},$$
$$[a, b] = [0, \infty), \quad p(x) = x^{s}\exp(-x).$$

Si la fonction à intégrer dans

$$I(f) = \int_0^{\omega} f(x)\,dx$$

est bien approchée par des polynômes trigonométriques de période ω, on a la formule

$$I(f) \approx S_N(f) = \frac{\omega}{N} \sum_{j=0}^{N-1} f\left(\frac{j\omega}{N}\right). \tag{16}$$

On a

$$S_N\left(\exp\left(2\pi m i \frac{x}{\omega}\right)\right) = \frac{\omega}{N} \sum_{j=0}^{N-1} \exp\left(2\pi m i \frac{j\omega}{N\omega}\right) =$$

$$= \begin{cases} \dfrac{\omega}{N} \displaystyle\sum_{j=0}^{N-1} 1 = \omega \text{ pour } \frac{m}{N} \text{ entier}, \\ \dfrac{\exp(2\pi m i) - 1}{\exp\left(\frac{2\pi m i}{N}\right) - 1} = 0 \text{ pour } \frac{m}{N} \text{ non entier}. \end{cases}$$

D'autre part,

$$I\left(\exp\left(2\pi m i\frac{x}{\omega}\right)\right)=\begin{cases}\omega \text{ pour } m=0,\\ 0 \text{ pour } m\neq 0.\end{cases}$$

Il s'ensuit que la formule (16) vérifie les fonctions $\cos\left(2\pi m\frac{x}{\omega}\right)$ pour $m=0$ ou pour $\frac{m}{N}$ non entier et toutes les fonctions $\sin\left(2\pi m\frac{x}{\omega}\right)$; elle est donc exacte pour tout polynôme trigonométrique

$$t_N(x)=a_0+\sum_{m=1}^{N-1}\left(a_m\cos\left(2\pi m\frac{x}{\omega}\right)+b_m\sin\left(2\pi m\frac{x}{\omega}\right)\right)+$$

$$+b_N\sin\left(2\pi N\frac{x}{\omega}\right),$$

par conséquent,

$$R_N(f)=R_N(f-t_N)+R_N(t_N)=R_N(f-t_N).$$

Pareillement à (8), on obtient l'évaluation

$$|R_N(f)|\leqslant 2\omega\inf_{t_N}\max_{[0,\,\omega]}|f(x)-t_N(x)|.$$

Les formules de quadrature (16) s'emploient tout particulièrement lorsqu'il s'agit de calculer des intégrales par rapport aux variables angulaires dans l'intégration en coordonnées sphériques et polaires.

On montre qu'il n'y a pas de formules de quadrature basée sur N points qui soient exactes pour tous les polynômes trigonométriques de degré N.

§ 4. Evaluation pratique de l'erreur dans des formules de quadrature élémentaires

Plus haut nous avons établi plusieurs formules de quadrature et évalué strictement leur erreur. Nous n'avons pas résolu pour autant tous les problèmes que nous pose l'intégration numérique. La tâche majeure du calcul mathématique est la création des algorithmes et de leurs systèmes permettant de résoudre les problèmes avec une précision donnée d'avance et un coût minimal en travail humain et machine. L'utilisation pratique des évaluations obtenues ci-dessus exige des calculs analytiques pour apprécier les dérivées, donc un grand effort de la part du chercheur; de plus, elles s'avèrent souvent trop fortes. Aussi pour mettre au point de pareils systèmes, on refuse en général de se servir de pareilles évaluations, sacrifiant souvent la petitesse garantie de l'erreur sur la solution approchée.

On peut dire qu'entre son apparition et sa résolution, tout problème traverse un système composé des personnes chargées de le résoudre et de la machine à calculer. A l'aube de l'histoire des calculatrices,

le goulot d'étranglement de ce système était le manque de machines. Ce qui justifie l'emploi de méthodes analytiques de résolution ou d'évaluation analytique des erreurs.

Depuis les machines à calculer se sont implantées dans les domaines les plus divers des activités humaines. Maintenant, le système en question souffre d'un choix trop lent du modèle mathématique, de la méthode de résolution, du programme et d'autres étapes antérieures à la résolution directe sur machine. Ces étapes se ralentissent encore plus si le calcul automatique est fait par des non-mathématiciens, disons des linguistes, médecins, économistes, géographes, peu versés dans les méthodes numériques ou la programmation. Leur initiation aux arcanes du métier risque de devenir une fin en soi qui leur fait oublier les problèmes fondamentaux de leur discipline et coûte assez cher à la société. Aussi est-il urgent aujourd'hui de créer des systèmes très faciles à manier qui n'exigent pas une connaissance poussée des méthodes numériques et de la programmation.

On exige tout naturellement, par exemple, que le programme de calcul d'une intégrale avec une précision donnée soit à la taille d'un chercheur qui, s'il sait ce qu'est l'intégrale, ne sait ni intégrer ni dériver.

D'autre part, la contribution des mathématiques à d'autres sciences n'est pas tant d'effectuer des simulations sur calculateurs que d'élaborer des modèles mathématiques. Le spécialiste qui élabore le modèle doit donc posséder une certaine culture mathématique.

Le but de nos raisonnements est de vous rappeler une chose évidente. L'efficacité de notre travail doit se mesurer non seulement au coût total de résolution de tels problèmes, mais aussi au préjudice occasionné à la société du fait que nous en avons laissé de côté certains autres, plus importants peut-être.

En analysant pratiquement l'erreur commise dans l'intégration numérique on se sert souvent de procédés semi-empiriques dont le plus usité est le suivant. On effectue les calculs par deux formules de quadrature

$$I(f)=\int_a^b f(x)\,dx \approx S^k(f)=\frac{b-a}{2}\sum_{j=1}^{N_k} D_j^k f(x_j^k),\quad k=1,2;$$

une certaine combinaison linéaire

$$S(f) = S^1(f) + \theta\,(S^2(f) - S^1(f))$$

de ces formules est prise pour valeur approchée de l'intégrale et $\rho = |S^2(f) - S^1(f)|$ pour mesure de l'erreur de la formule approchée $I(f) \approx S(f)$. $\theta = 0$ représente un cas assez typique.

Le procédé ci-dessus n'étant pas univoque, on ne saurait le justifier complètement.

Soit $S^1(f)$ la formule de Simpson :

$$\int_{-1}^{1} f(x)\,dx \approx S^1(f) = \frac{f(-1)+4f(0)+f(1)}{3},$$

et $S^2(f)$ celle des trapèzes : $S^2(f) = f(-1) + f(1)$ et $\theta = 0$. L'erreur est alors mesurée par

$$\rho = \frac{2}{3}\,|f(1) - 2f(0) + f(-1)|.$$

Si $S^2(f) = 2f(0)$ est la formule des rectangles, on a respectivement $\rho = \frac{1}{3}\,|f(1) - 2f(0) + f(-1)|$. Nous avons donc obtenu deux estimations empiriques de l'erreur dans une même formule de Simpson.

Essayons de voir plus clair. $S(f)$ est une certaine somme de quadrature

$$S(f) = \frac{b-a}{2} \sum_{j=1}^{N} D_j f(x_j) \tag{1}$$

sur l'ensemble des points x_j appartenant à la réunion des points correspondant à $S^1(f)$ et $S^2(f)$. On a par ailleurs

$$\rho = |\,l(f)\,|,$$

où

$$l(f) = \frac{b-a}{2} \sum_{j=1}^{N} B_j f(x_j). \tag{2}$$

Prenons une combinaison linéaire de la forme (2) et posons

$$S^1(f) = S(f), \quad S^2(f) = S(f) - \theta l(f).$$

La valeur approchée obtenue de l'intégrale $I(f) \approx S(f)$ est entachée de l'erreur $\rho = \theta\,|\,l(f)\,|$. Nous voyons qu'en poursuivant dans cette voie on peut obtenir une infinité d'estimations de l'erreur sur une même formule de quadrature (1).

Le problème considéré peut s'énoncer comme suit. La formule

$$I(f) \approx S(f) = \frac{b-a}{2} \sum_{j=1}^{N} D_j f(x_j) \tag{3}$$

permet de calculer la valeur approchée de l'intégrale. On désire former une expression du type (2) donnant une idée de l'erreur provenant de la formule de quadrature (3).

Supposons que cette erreur soit

$$R(f) = \overline{D}\,(b-a)^{m+1} f^{(m)}(\zeta). \tag{4}$$

Considérons le cas $m < N$. Pour $l(f)$ on peut prendre

$$l(f) = \overline{D}(b-a)^{m+1} m! f(x_{i_1}, \ldots, x_{i_{m+1}}),$$

où $x_{i_1}, \ldots, x_{i_{m+1}}$ sont des points différents de la formule de quadrature (3). L'expression de la différence divisée en fonction de la dérivée entraîne

$$l(f) = \overline{D}(b-a)^{m+1} f^{(m)}(\zeta), \quad a < \zeta < b;$$

donc, si $a = \text{const}$, $f^{(m)}(a) \neq 0$, $b - a \to 0$, on a $R(f) \sim l(f)$, et $\rho = |l(f)|$ mesure l'erreur commise. Supposons par exemple qu'on évalue l'erreur résultant de la formule des trapèzes

$$\int_a^b f(x)\,dx \approx S(f) = \frac{b-a}{2}\left(f(a) + 0 \cdot f\left(\frac{a+b}{2}\right) + f(b)\right);$$

suivant les estimations des §§ 1, 2, on a

$$R(f) = -\frac{(b-a)^3}{12} f''(\zeta);$$

ainsi, on prend pour mesure de l'erreur

$$\frac{b-a}{3}\left|f(b) - 2f\left(\frac{a+b}{2}\right) + f(a)\right|.$$

Il en va autrement avec $m \geqslant N$. Les quantités $f(x_1), \ldots, f(x_N)$ ne permettent alors aucune approximation de $f^{(m)}(\zeta)$ et le problème d'évaluation empirique de l'erreur tel qu'il y a été posé plus haut est insoluble. Il est impossible par exemple de se faire une idée satisfaisante de l'estimation de l'erreur dans la formule de Simpson à l'aide des valeurs $f(a)$, $f\left(\frac{a+b}{2}\right)$, $f(b)$, encore qu'une surestimation soit possible.

Voici une approche du problème surgi. Examinons la formule (2.5) pour $n = N - 1$. L'erreur est alors évaluée par

$$|R(f)| \leqslant (b-a)^N \overline{D} \max_{[a,\,b]} |f^{N-1}(x)| = \sigma.$$

Posons

$$\rho = (b-a)^N \overline{D}\,(N-1)!\,|f(x_1; \ldots; x_N)|. \tag{5}$$

Si $a = \text{const}$, $f^{(N-1)}(a) \neq 0$, $b - a \to 0$, on a $\sigma \sim \rho$ et ρ peut être prise pour estimation approchée de l'erreur dans (3). C'est une estimation trop forte puisque l'hypothèse $m \geqslant N$ entraîne $R(f) = O((b-a)^{N+1})$; nous n'en disposons pourtant pas de meilleure par rapport à l'évaluation par σ. Pour la formule de Simpson,

(2.5) donne (5) où $\overline{D} = \frac{1}{81}$, et l'erreur est donc mesurée par

$$\frac{4(b-a)}{81}\left|f(b)-2f\left(\frac{a+b}{2}\right)+f(a)\right|. \tag{6}$$

S'il s'agit de calculer l'intégrale d'une fonction analytique et si le nombre de points est donné, les formules de Gauss paraissent les mieux indiquées car elles sont exactes pour des polynômes de degré maximal. On évalue l'erreur suivant le schéma décrit.

Nous avons proposé d'évaluer l'erreur sur la formule de Simpson à l'aide de l'expression (6) en supposant $f(x)$ suffisamment régulière. Néanmoins, de pareilles conclusions empiriques sont souvent appliquées sans que les conditions imposées à leur obtention soient satisfaites, par exemple dans l'intégration des fonctions non régulières.

Soit

$$[a, b]=[0, 1], \quad f_{\alpha,\beta}(x)=(x+\alpha)^{\beta}, \quad \alpha>0, \ \beta>0.$$

Alors la formule de Simpson est affectée de l'erreur

$$E_{\alpha,\beta}=\frac{(1+\alpha)^{\beta+1}-\alpha^{\beta+1}}{\beta+1}-\left(\frac{1}{6}\alpha^{\beta}+\frac{2}{3}\left(\frac{1}{2}+\alpha\right)^{\beta}+\frac{1}{6}(1+\alpha)^{\beta}\right).$$

D'autre part,

$$\left|f_{\alpha,\beta}(1)-2f_{\alpha,\beta}\left(\frac{1}{2}\right)+f_{\alpha,\beta}(0)\right|=S_{\alpha,\beta}=\left|(1+\alpha)^{\beta}-2\left(\frac{1}{2}+\alpha\right)^{\beta}+\alpha^{\beta}\right|.$$

On a

$$\lim_{\beta\to 0}\lim_{\alpha\to 0}\frac{E_{\alpha,\beta}}{S_{\alpha,\beta}}=\frac{1}{6}.$$

Ainsi, l'emploi de (6) en vue d'évaluer l'erreur dans la formule de Simpson avec α et β petits est injustifié.

Si α et β sont faibles, l'estimation

$$\frac{1}{6}(b-a)\left|f(b)-2f\left(\frac{a+b}{2}\right)+f(a)\right| \tag{7}$$

paraît plus naturelle.

Dans un programme standard d'intégration où l'on a essayé d'obtenir le résultat avec le « maximum » de sûreté on mesurait l'erreur par

$$D^0(b-a)^N\,|f(x_1;\ \ldots;\ x_N)|$$

avec D^0 ainsi défini : on a pris le segment [0, 1] et posé

$$D^0=\sup_{\substack{0<\alpha\\ 0<\beta\leqslant N-1}}\frac{|R(f_{\alpha,\beta})|}{|f_{\alpha,\beta}(x_1;\ \ldots;\ x_N)|};$$

pour la formule de Simpson un tel choix de D^0 conduit aussi à évaluer l'erreur par (7).

Dans le cas d'intégrales multiples tous les procédés usuels d'évaluation de l'erreur partent du schéma initial que nous venons de critiquer. C'est que dans

le cas multidimensionnel l'erreur est évaluée moyennant les valeurs de plusieurs dérivées de la fonction à intégrer. L'obtention d'estimations « justifiées » analogues à (6) paraît ici une affaire très ardue. Aussi s'adresse-t-on au schéma initial dont on vérifie ensuite les résultats. Dans le calcul de l'intégrale par la formule

$$I(f) \approx \sum_{j=1}^{N} D_j f(P_j)$$

dont l'erreur est évaluée par $|\, l(f)\,|$, $l(f) = \sum_{j=1}^{N} B_j f(P_j) \not\equiv 0$ les paramètres libres sont les quantités D_j, B_j et les coordonnées des points de base P_j de la formule de quadrature. Le choix de ces paramètres se fait selon la règle suivante. On se donne le degré m des polynômes vérifiant cette formule et on suppose que $l(f)$ s'annule sur des polynômes d'un certain degré m_1 donné d'avance ($m_1 < m$) et ne possède pas cette propriété pour tous les polynômes de degré m ; on pose d'ordinaire $m_1 = m-2$. Par un choix convenable de B_j, D_j et P_j on prend un N minimal satisfaisant aux conditions imposées. Dans ces conditions $l(f)$ est définie de façon non univoque à un certain facteur α près : $l(f) = \alpha l^0(f)$ avec $l^0(f) = \sum_{j=1}^{N} B_j^0 f(P_j)$. Généralement α est égalé au D_j/B_j^0 de plus petit module.

§ 5. Intégration de fonctions fortement oscillantes

Soit à calculer

$$\int_a^b f(x) \exp(i\omega x)\, dx,$$

où $\omega(b-a) \gg 1$, $f(x)$ est une fonction régulière. Les fonctions

$$\mathrm{Re}\,(f(x)\exp(i\omega x)), \quad \mathrm{Im}\,(f(x)\exp(i\omega x))$$

admettent sur le segment considéré environ $\omega(b-a)/\pi$ zéros. Puisqu'un polynôme de degré n possède au plus n zéros, ces fonctions ne sont bien approchées par des polynômes de degré n que pour $n \gg \omega(b-a)/\pi$. Si l'on veut donc calculer directement les intégrales de telles fonctions, on utilisera des formules de quadrature exactes pour des polynômes de degré supérieur.

On gagnerait en considérant la fonction $\exp(i\omega x)$ comme une fonction de poids.

Comme au § 1, donnons-nous les points d'interpolation

$$x_j = \frac{b+a}{2} + \frac{b-a}{2} d_j, \quad j = 1, \ldots, n,$$

et remplaçons l'intégrale de départ par $\int_a^b L_n(x) \exp(i\omega x)\, dx$; celle-ci s'explicite en

$$\int_a^b L_n \exp(i\omega x)\, dx = S_n^{\omega}(f) =$$

$$= \frac{b-a}{2} \exp\left(i\omega \frac{a+b}{2}\right) \sum_{j=1}^{n} D_j\left(\omega \frac{b-a}{2}\right) f(x_j),$$

où

$$D_j(p) = \int_{-1}^{1} \left(\prod_{k \neq j} \frac{t-d_k}{d_j-d_k} \right) \exp(ipt)\, dt. \tag{1}$$

On aboutit à la formule

$$\int_a^b f(x) \exp(i\omega x)\, dx \approx S_n^{\omega}(f) \tag{2}$$

avec pour reste

$$R_n(f) = \int_a^b (f(x) - L_n(x)) \exp(i\omega x)\, dx.$$

Selon (1.1)

$$|R_n(f)| \leqslant \int_a^b |f(x) - L_n(x)|\, dx \leqslant$$

$$\leqslant D(d_1, \ldots, d_n) \max_{[a,\,b]} |f^n(x)| \left(\frac{b-a}{2}\right)^{n+1}.$$

Le calcul de telles intégrales est un problème type qu'on rencontre en développant des fonctions en série de Fourier, en construisant les diagrammes de directivité d'antennes, etc.

Les programmes standard de calcul utilisent les formules (1), (2) correspondant à $n = 3$, $d_1 = -1$, $d_2 = 0$, $d_3 = 1$ (*formule de Filon*) ou à $n = 5$, $d_1 = -1$, $d_2 = -0{,}5$, $d_3 = 0$, $d_4 = 0{,}5$, $d_5 = 1$.

Le calcul suivant les formules (1), (2) peut être cause d'une difficulté que nous allons illustrer pour le cas $n = 2$, $d_1 = -1$, $d_2 = 1$; on a alors

$$D_1(p) = \int_{-1}^{1} \frac{1-t}{2} \exp(ipt)\, dt = \frac{\sin p}{p} + \frac{p \cos p - \sin p}{p^2}\, i,$$

$$D_2(p) = \int_{-1}^{1} \frac{1+t}{2} \exp(ipt)\, dt = \frac{\sin p}{p} - \frac{p \cos p - \sin p}{p^2}\, i.$$

Si $p \to 0$, on a

$$\frac{p\cos p - \sin p}{p^2} = -\frac{p}{3} + O(p^3) \to 0.$$

Donc, $D_1(p)$, $D_2(p) \to 1$ pour p tendant vers zéro.

Supposons p petit. Les fonctions $\cos p$ et $\sin p$ sont calculées sur machine avec une erreur $O(2^{-t})$ et $O(p2^{-t})$ respectivement, et les coefficients $D_1(p)$ et $D_2(p)$ acquièrent une erreur $O(2^{-t}/p)$. Si $n > 2$, l'erreur sur les coefficients $D_j(p)$ calculés par les formules (1) peut être de l'ordre de $2^{-t}/p^{n-1}$. Par exemple, pour $t = 30$, $n = 5$, $p = 0{,}01$, une telle erreur n'est plus permise. Aussi les programmes de calcul d'intégrales de fonctions oscillantes doivent-ils posséder un bloc destiné à modifier les formules numériques pour p petits afin d'éviter une erreur de calcul essentielle. Si n n'est pas très grand, disons $n = 2$, la disposition des calculs peut être la suivante : pour $|p|$ supérieur à un certain p_n choisi expérimentalement les calculs se font selon les formules (1), (2) ; pour $|p| \leqslant p_n$ on se sert, pour trouver l'intégrale de départ, de la formule des trapèzes (1.7) et on prend toute la fonction $f(x)\exp(i\omega x)$ comme fonction à intégrer. Pour notre cas ($n = 2$, $d_1 = -1$, $d_2 = 1$), la formule (1.7) s'écrit

$$\int_a^b f(x)\exp(i\omega x)\,dx \approx \frac{b-a}{2}(\exp(i\omega a) f(a) + \exp(i\omega b) f(b)). \quad (3)$$

On regroupe (1) et (3) en une seule formule

$$\int_a^b f(x)\exp(i\omega x)\,dx = \frac{b-a}{2}\exp\left(i\omega\frac{a+b}{2}\right)(D_1(p) f(a) + D_2(p) f(b)) \quad (4)$$

avec

$$p = \omega\frac{b-a}{2},$$

$$D_{1,2}(p) = \begin{cases} \dfrac{\sin p}{p} \pm \dfrac{p\cos p - \sin p}{p^2} i & \text{pour } |p| > p_2, \\ \exp(\mp pi) & \text{pour } |p| \leqslant p_2. \end{cases} \quad (5)$$

Les programmes standard de calcul d'intégrales du type considéré font appel aux formules de quadrature de la forme (4), (5). Une question se pose tout naturellement : pourquoi cette complication des programmes ? On ferait peut-être mieux de rédiger un programme standard de calcul par les formules (2), un autre par la formule des trapèzes et de lancer l'instruction suivante : pour $|p|$ grands appeler le premier programme, pour $|p|$ petits le second. Voyons si cette politique est rationnelle. Notre intention est de faciliter au

maximum le travail de l'utilisateur. Si la description d'un programme standard est trop détaillée, l'utilisateur peut s'embrouiller et

1) appeler un autre programme de qualité inférieure mais mieux décrit et plus facile à manipuler;

2) exécuter mal le programme et n'obtenir aucun résultat; dans notre cas, par exemple, la méthode (1), (2) pour $p = 0$ conduit à l'arrêt de l'ordinateur;

3) exploiter mal le programme et aboutir à un résultat erroné en le supposant juste, ce qui arrivera si par exemple l'on se sert des formules (1), (2) avec $2^{-t}/p^n$ de l'ordre de l'unité.

Il paraît que le cas 3) est le plus désastreux.

En choisissant la méthode, on se souciera que la description correspondante du programme soit simple et concise à souhait, car tout renseignement supplémentaire peut recevoir une interprétation incorrecte et limiter le domaine d'application du programme. L'utilisation de deux programmes, l'un pour $|p| > p_2$, l'autre pour $|p| \leqslant p_2$, n'a fait que compliquer les choses, surtout dans le cas des séries d'intégrales dont certaines sont à p petits et d'autres à p grands. On leur a donc préféré un seul programme qui effectue les calculs suivant les formules (4), (5) et dont la description ne mentionne pas la quantité $|p|$.

Voici un exemple instructif.

La description d'un programme standard de calcul d'intégrales multiples stipulait entre autres: « Le présent programme n'est pas à recommander pour un nombre de points supérieur à 100 000. » Peu après, ce programme a été retiré de la circulation. Il s'est avéré que dans leur grande majorité les premiers utilisateurs se donnaient un n égal à 100 000, et le calcul d'une intégrale simple prenait trop de temps machine. La nouvelle s'est très vite répandue et les calculateurs ont vite abandonné le programme en question. Or, celui-ci permettait en fait de calculer la plupart des intégrales réelles avec une précision acceptable si le nombre de points était de l'ordre de 1000; quant au nombre 100 000, il représentait simplement une borne supérieure approximative des valeurs de n pour lesquelles l'influence de l'erreur de calcul sur le résultat n'était pas encore catastrophique.

§ 6. Amélioration de la précision d'intégration par partage équidistant du segment

L'évaluation (3.4) entraîne que l'erreur dans la formule de quadrature est évaluée au moyen de la précision d'approximation de la fonction $f(x)$ par des polynômes de degré n. La tentative d'améliorer la précision en élevant le degré des polynômes vérifiant cette formule semble donc parfaitement naturelle. Cette voie est cependant semée d'écueils. Avec un choix malheureux de points il peut arriver que $\sum_{j=1}^{n} |D_j|$ augmente avec n. V_n dans (3.4) croît donc lui aussi avec n et la diminution de $E_m(f)$ pour un n croissant peut ne pas compenser l'augmentation de V_n.

Par exemple, pour une répartition uniforme élémentaire des points

$$d_j^{(n)} = -1 + \frac{2(j-1)}{n-1}, \quad j = 1, \ \ldots, \ n, \ [a, b] = [-1, 1],$$

V_n augmente approximativement comme 2^n (plus précisément, $\log_2 V_n \sim n$) si $m = n - 1$. Il en résulte que la fonction analytique $f(x) = (1 + 25x^2)^{-1}$ présente pour de tels points une erreur

$$R_n(f) = \int_{-1}^{1} f(x)\,dx - S_n(f) \to \infty,$$

$n \to \infty$.

Prenons le segment d'intégration $[-1, 1]$ et énonçons un théorème général recommandant la prudence dans l'usage des formules exactes pour des polynômes de très haut degré. Soit pour tout n la formule

$$\int_{-1}^{1} f(x)\,dx \approx \sum_{j=1}^{n} D_j^{(n)} f(d_j^{(n)}), \tag{1}$$

stricte pour les polynômes de degré $n - 1$. Notons $w_n(x_1, x_2)$ le nombre de points de (1) appartenant au segment $[x_1, x_2]$.

THÉORÈME (sans démonstration). *Soit un segment* $[x_1, x_2] \in [-1, 1]$ *tel que*

$$\frac{w_n(x_1, x_2)}{n} \not\to \int_{x_1}^{x_2} \frac{1}{\pi\sqrt{1-x^2}}\,dx \quad \textit{pour} \quad n \to \infty.$$

Alors on peut indiquer un $b \neq 0$ *et un* a *tels que l'une des fonctions analytiques vérifie*

$$\operatorname{Re}\left(\frac{1}{x-(a+bi)}\right), \quad \operatorname{Im}\left(\frac{1}{x-(a+bi)}\right),$$

$$\overline{\lim_{n\to\infty}} \, |R_n(f)| = \infty.$$

Ainsi, la densité de répartition des points des formules de quadrature (1) exactes pour les polynômes de degré $n - 1$ doit être égale à celle des zéros des polynômes orthogonaux, i.e. des zéros des formules de Gauss. Dans le cas contraire ces formules ne peuvent pas être considérées comme universelles.

Récrivons l'évaluation (3.8) de l'erreur dans les formules de Gauss de façon suivante:

$$|R_n(f)| \leqslant 2\left(\int_a^b p(x)\,dx\right) E_{2n-1}(f). \tag{2}$$

Supposons que la fonction à intégrer $f(x)$ soit continue; le théorème de Weierstrass dit alors que pour tout $\varepsilon > 0$ il existe un polynôme $P_q(x)$ tel que $\max_{[a, b]} |f(x) - P_q(x)| \leqslant \varepsilon$, d'où $E_m(f) \to 0$ avec

$m \to \infty$. Par conséquent, pour n'importe quelle fonction $f(x)$ continue, l'erreur commise dans les formules de Gauss tend vers 0 pour $n \to \infty$.

PROBLÈME. Soit $f(x)$ une fonction intégrable au sens de Riemann. Démontrer que pour les formules de Gauss on a

$$R_n(f) \to 0 \quad \text{pour } n \to \infty.$$

On voit que les programmes universels de calcul d'intégrales des fonctions pourraient très bien se baser sur les formules de quadrature de Gauss. Il y a cependant un inconvénient: il est impossible de prédire le nombre de points des formules de quadrature dans le calcul d'intégrales avec une précision donnée. Aussi faut-il introduire dans la machine ou enregistrer en mémoire rapide à partir de mémoires auxiliaires les points et les poids des formules qui correspondent à un n de plus en plus grand, ce qui ralentit quelque peu les calculs, notamment sur de petites machines. Signalons de plus qu'on voit souvent surgir le problème de calcul d'intégrales telles que la fonction à intégrer ou ses dérivées d'ordre peu élevé présentent des intervalles de brusque variation, par exemple deviennent infinies. De telles fonctions se prêtent mal à l'approximation par des polynômes sur le segment d'intégration tout entier. Il y a donc souvent intérêt à partager le segment initial en parties, puis à appliquer à chaque segment partiel les formules de Gauss ou autres.

Soit à trouver l'intégrale $I = \int_A^B f(x)\,dx$. Divisons $[A, B]$ en M parties $[a_{q-1}, a_q]$ avec $a_0 = A$, $a_M = B$. Pour calculer l'intégrale sur chacune des parties, servons-nous de l'une quelconque des formules de quadrature des §§ 1, 3:

$$I_q = \int_{a_{q-1}}^{a_q} f(x)\,dx \approx s_q = \\ = \frac{a_q - a_{q-1}}{2} \sum_{j=1}^{m} D_j f\left(\frac{a_{q-1}+a_q}{2} + \frac{a_q - a_{q-1}}{2} d_j\right) \qquad (3)$$

avec

$$D\left(\max_{[a_{q-1},\, a_q]} |f^{(n)}(x)|\right)\left(\frac{a_q - a_{q-1}}{2}\right)^{n+1}$$

pour évaluation du reste. L'intégrale sur le segment tout entier sera alors approchée par la somme

$$S_M^m(f) = \sum_{q=1}^{M} s_q = \sum_{q=1}^{M} \frac{a_q - a_{q-1}}{2} \sum_{j=1}^{m} D_j f\left(\frac{a_{q-1}+a_q}{2} + \frac{a_q - a_{q-1}}{2} d_j\right) \qquad (4)$$

dont le reste est évalué par

$$|R_M^m(f)| = |I - S_M^m(f)| \leqslant D \sum_{q=1}^{M} \left(\max_{[a_{q-1}, a_q]} |f^{(n)}(x)|\right) \left(\frac{a_q - a_{q-1}}{2}\right)^{n+1}. \quad (5)$$

On dit de la formule (4) qu'elle est *composite* ou *généralisée*.

Examinons un cas dont l'étude est particulièrement simple, à savoir celui de segments partiels égaux: $a_q - a_{q-1} \equiv H$.

Après avoir remplacé $\max\limits_{[a_{q-1}, a_q]} |f^{(n)}(x)|$ par $A_n = \max\limits_{[a, b]} |f^{(n)}(x)|$ l'estimation d'erreur (5) devient alors

$$|R_M^m(f)| \leqslant \bar{D} A_n (B - A) H^n, \quad \text{où} \quad \bar{D} = D 2^{-(n+1)}, \quad (6)$$

ou

$$|R_M^m(f)| \leqslant \bar{D} A_n \frac{(B-A)^{n+1}}{M^n}. \quad (7)$$

Ne pensez pas que pour des fonctions possédant des dérivées bornées peu nombreuses, les formules composites d'intégration numérique convergent mieux que les formules de Gauss. Supposons que la fonction à intégrer possède p dérivées bornées. Par application de la formule (4) correspondant à $n = p$, on obtient une valeur approchée entachée d'une erreur $O(N^{-p})$. On sait par ailleurs que pour une telle fonction $E_{2N-1}(f) = O(N^{-p})$. La formule (3.8) entraîne donc l'estimation suivante de l'erreur dans la formule de Gauss à N points:

$$R_N(f) = O(N^{-p}).$$

L'ordre de l'estimation d'erreur est le même dans les deux cas.

Les formules de Gauss appliquées à tout le segment d'intégration présentent un autre avantage. On n'a pas à évaluer le nombre p_0 de dérivées bornées de la fonction à intégrer ni à choisir, par conséquent, la formule la plus indiquée d'intégration numérique sur les segments partiels, car les formules de Gauss garantissent automatiquement une erreur de l'ordre de $O(N^{-p})$. Si les dérivées d'un ordre supérieur à p_0 ne deviennent infinies qu'aux extrémités du segment, l'ordre de convergence des formules de Gauss est d'ordinaire plus élevé. On conçoit qu'une formule convergeant plus rapidement n'est pas forcément plus précise qu'une autre dont l'ordre de convergence est plus petit (cf. les explications des fig. 3.15.2-3.15.5).

Donnons certaines formules de quadrature et évaluons les erreurs dans les cas particuliers des formules (3).

1. *Formule des trapèzes généralisée à pas d'intégration constant.* Soit le pas constant $a_q - a_{q-1} \equiv H$. La formule (3) devient

$$\int_{a_0}^{a_M} f(x)\,dx \approx H\left(\frac{f(a_0)}{2} + f(a_1) + \ldots + f(a_{M-1}) + \frac{f(a_M)}{2}\right),$$

et le reste est évalué par

$$A_2 \frac{(a_M - a_0) H^2}{12} = A_2 \frac{(a_M - a_0)^3}{12M^2}. \quad (8)$$

2. *Formule de Simpson généralisée à pas d'intégration constant.* Pour le pas constant $a_q - a_{q-1} \equiv H = 2h$, la formule (3) se transforme en

$$\int_{a_0}^{a_M} f(x)\,dx \approx H\left(\frac{1}{6}f(a_0)+\frac{4}{6}f(a_{\frac{1}{2}})+\frac{1}{3}f(a_1)+\right.$$
$$\left.+\frac{4}{6}f(a_{\frac{3}{2}})+\frac{1}{3}f(a_2)+\ldots+\frac{1}{3}f(a_{M-1})+\frac{4}{6}f(a_{M-\frac{1}{2}})+\frac{1}{6}f(a_M)\right),$$

où $a_j = a_0 + jH$. L'estimation du reste s'écrit

$$A_4\frac{(a_M-a_0)\,H^4}{2880} = \frac{A_4\,(a_M-a_0)^5}{2880M^4} = \frac{A_4\,(a_M-a_0)\,h^4}{180}.$$

Cette dernière écriture est la plus usitée.

Nous venons d'obtenir les formules avec une erreur en $O(N^{-n})$ pour un nombre total $N = Mm$ de points d'intégration sous l'hypothèse des $|f^{(n)}(x)|$ bornées. Les formules des trapèzes et de Simpson sont basées sur un nombre N de points inférieur à Mm, puisque les extrémités des segments élémentaires $[a_{q-1}, a_q]$ étaient les points d'interpolation et les valeurs de la fonction en ces extrémités servaient à calculer les intégrales sur deux segments élémentaires.

On aboutit à l'estimation du même ordre dans l'hypothèse plus faible $\int_A^B |f^{(n)}(x)|\,dx < \infty$.

En effet, rappelons-nous la représentation (2.3) de l'erreur sur un segment élémentaire

$$R_N(f) = \int_a^b K_N(y)\,f^{(n)}(y)\,dy,$$

où

$$K_N(y) = \frac{(b-y)^n}{n!} - (b-a)\sum_{j=1}^{N} D_j\,\frac{\overline{(x_j-y)}^{n-1}}{(n-1)!}. \tag{9}$$

D'où l'évaluation

$$|R_N(f)| \leqslant \max_{[a,\,b]} |K_N(y)| \int_a^b |f^{(n)}(y)|\,dy. \tag{10}$$

Par ailleurs, on a

$$\max_{[a,\,b]} |K_N(y)| = \widetilde{D}\,(b-a)^n$$

avec

$$\widetilde{D} = \max_{[-1,\,1]}\left|\frac{(1-t)^n}{n!} - \sum_{j=1}^{N} D_j\,\frac{\overline{(d_j^0-t)}^{n-1}}{(n-1)!}\right|, \quad d_j^0 = \frac{x_j-a}{b-a}.$$

Cette égalité est obtenue par changement de variable $y = a + (b-a)t$ dans le second membre de (10). Finalement

$$|R_N(f)| \leqslant \widetilde{D}(b-a)^n \int_a^b |f^{(n)}(x)|\,dx. \tag{11}$$

Ainsi, le reste de la formule de quadrature (3) est également évalué par

$$\widetilde{D}(a_q - a_{q-1})^n \int_{a_{q-1}}^{a_q} |f^{(n)}(x)|\,dx.$$

Soit $a_q - a_{q-1} \equiv H$. L'estimation totale du reste est alors

$$|R_M^m(f)| \leqslant \widetilde{D} \sum_{q=1}^{M} H^n \int_{a_{q-1}}^{a_q} |f^{(n)}(x)|\,dx = \widetilde{D}H^n \int_A^B |f^{(n)}(x)|\,dx. \tag{12}$$

On demande, par exemple, de trouver

$$\int_0^1 (\sin x)^\alpha\,dx, \quad 0 < \alpha < 1.$$

Comme $((\sin x)^\alpha)'_x \to \infty$ pour $x \to 0$, il est impossible d'évaluer l'erreur à l'aide de $\max |f'(x)|$. D'autre part, $(\sin x)^\alpha$ est monotone sur $[0, 1]$, donc

$$\int_0^1 |((\sin x)^\alpha)'|\,dx = (\sin x)^\alpha \Big|_0^1 = 1.$$

Par conséquent, avec la formule de quadrature (4) correspondant à $n = 1$ à pas constant $a_q - a_{q-1} \equiv H$, on obtient, conformément à (12), une estimation d'erreur de l'ordre de $O(H)$.

Dans le présent exemple la petitesse de l'erreur ne découlait pas de l'estimation (6) ; d'autre part, il résulte de (12) qu'il s'agit d'une erreur en $O(1/N)$. C'est pourquoi une estimation importante de l'erreur ne signifie nullement que celle-ci est considérable.

En particulier, on améliore fortement l'évaluation de l'erreur dans le cas multidimensionnel si l'on utilise les normes des dérivées dans les métriques L_p, $p \neq \infty$.

On se demande alors si, dans les mêmes hypothèses sur la fonction à intégrer, on peut obtenir des évaluations d'erreur plus précises ou au moins améliorer les constantes y figurant. Ceci nous conduit à étudier la construction de formules de quadrature avec une évaluation optimale de l'erreur, c'est-à-dire de *formules de quadrature optimales*.

§ 7. Position des problèmes d'optimisation

Nous avons construit plusieurs formules d'intégration numérique et nous pourrions continuer dans cette voie. Un problème surgit alors : plus les procédés de résolution sont nombreux et plus on hésite à fixer son choix, car chacun possède peut-être des avantages tels que la facilité de programmation, un temps machine réduit, un faible encombrement de mémoire, une évaluation d'erreur simple, de vastes débouchés. Si bien qu'une information trop abondante sur les techniques de résolution, et qui augmente démesurément, peut être cause du surplace. Une systématisation des méthodes, leur sélection s'imposent donc. On voudrait, cela va de soi, résoudre son problème au mieux, par un procédé optimal. Nous avons déjà considéré certaines formulations types des problèmes afférents aux méthodes optimales de résolution en calculant les valeurs de fonctions ; les problèmes à résoudre se prêtent alors à des positions mathématiques déterminées.

Dès qu'un problème surgit, on se propose d'en donner la description la plus détaillée possible, puis de lui trouver la meilleure solution. Comme on n'y réussit généralement pas, on se contente soit de le décrire au mieux et de le résoudre de façon satisfaisante, soit, au contraire, de le décrire de facon satisfaisante pour, ensuite, le résoudre complètement. En ce qui concerne l'optimisation des méthodes, il s'agit du choix entre une résolution satisfaisante pour des classes de problèmes proches des problèmes réels et une résolution complète pour des classes mathématiques types semblables à celle qui sera étudiée dans le paragraphe suivant. Mais pourquoi parler de la « résolution satisfaisante d'un problème d'optimisation de méthodes » pour telles classes? Celui-ci ne se réduit-il pas à un problème mathématique bien concret qui est susceptible d'être entièrement résolu et pas seulement « de façon satisfaisante »?

C'est juste en principe, mais dans la pratique on n'arrive pas à résoudre entièrement un problème dans un temps acceptable: d'habitude, la construction d'une méthode optimale coûte sensiblement plus en temps que l'élaboration d'une description nouvelle, améliorée, des classes de problèmes considérés. Il ne faut pas oublier non plus que cette description mathématique n'admet pas toujours une formalisation : on dispose d'une représentation qualitative de la classe et de méthodes numériques valables, et l'on sent intuitivement que les méthodes sont à l'optimum sur le plan pratique ; or, l'on ne dispose même pas d'une description formelle nette de la classe des problèmes à résoudre. Un exemple de cet état de choses est l'intégration des fonctions régulières par morceaux que nous allons mettre à l'étude au § 17 (sous forme non formelle). De même, dans l'analyse des modèles physiques, ou bien on décrit le problème à l'aide d'un système compliqué d'équations différentielles et on

résout celui-ci avec une précision faible ou bien on le décrit par un système simple dont on trouve la solution avec une très grande précision.

Selon le contexte, on choisit telle ou autre approche. Il paraît qu'une science appliquée a souvent plus d'intérêt à poser des problèmes variationnels d'une façon correcte que de les résoudre entièrement. En général, une résolution satisfaisante de tel problème variationnel bien posé sur le plan pratique s'avère plus efficace qu'une résolution complète de tel autre ne rendant pas compte de toutes les particularités du phénomène.

Qu'est-ce qu'une méthode optimale de résolution d'un problème?

Certains entendent par résolution optimale une résolution qui coûte le moins en temps machine. C'est là une définition malheureuse car en principe tout problème peut être résolu sans faire appel à des calculatrices au prix d'un effort considérable.

D'autres disent que c'est la résolution la moins coûteuse en temps (ou en argent). (Notons qu'elle comprend déjà le coût de recherche de l'algorithme optimal.)

En voulant optimiser la résolution d'un problème nous ne tenons pas compte des moyens mis à la disposition du chercheur. Or, selon ces moyens (ordinateur, bibliothèque, personnel auxiliaire, collaborateurs rompus à la résolution de problèmes analogues, facultés du chercheur), la résolution optimale diffère d'une façon générale.

Nous n'avons parlé jusqu'à présent que de la résolution optimale de problèmes concrets. A vrai dire, cette approche n'est pas juste. Car une équipe de chercheurs a affaire à tout un ensemble de problèmes. Nous pouvons nous proposer de résoudre le premier, le second, etc., de ces problèmes en temps minimal. La situation changera radicalement si nous désirons d'en résoudre toute une série, et cela le moins cher possible. Nous résoudrons bien le premier, le deuxième, etc., problème par des méthodes optimales. Mais la science est en continuel progrès, et si nous n'étudions pas de nouveaux procédés, ne créons pas de nouveaux algorithmes et programmes standard à s'en servir à l'avenir, nous perdrons dans l'ensemble. En présence d'une grande série de problèmes pas très urgents, on ferait mieux de s'occuper de recherches théoriques, d'élaborer de nouveaux algorithmes, puis de s'attaquer à la résolution.

La question à savoir à qui confier la résolution optimale n'admet pas elle non plus de réponse univoque. S'il s'agit d'un seul problème concret n'exigeant a priori ni un algorithme inédit ni un temps machine considérable, on en chargera un utilisateur de moyenne qualification. Des problèmes compliqués en nombre important appellent des chercheurs hautement qualifiés qui donneront ici le meilleur d'eux-mêmes. On ne négligera certes pas d'étudier la pratique de la résolution de problèmes tant élémentaires qu'ardus. Pour qu'un problème d'optimisation des méthodes puisse être considéré comme

un problème purement mathématique, on doit déterminer la fonction économique (ou objectif), la classe de problèmes analysés et les possibilités du chercheur.

La qualité d'un algorithme sur une classe de problèmes est jugée sur le problème le plus difficilement maniable de cette classe. Donc, plus la classe est restreinte et plus puissants sont les algorithmes correspondants.

On pourrait penser qu'il faut construire des algorithmes optimaux pour le plus de classes possible. Or, en individualisant celles-ci outre mesure la recherche des procédés optimaux coûtera trop d'efforts, si bien que les forces et le temps pourront manquer pour résoudre les problèmes mêmes ainsi que pour élaborer et écrire les programmes usuels. Ceci concerne moins la résolution sur machines et dispositifs spécialisés, cas où il faut réduire au maximum la classe de problèmes étudiés et examiner la question sous tous ses aspects.

Il semble de prime abord que dans la pratique on a toujours affaire à des problèmes concrets, sans jamais envisager une classe de problèmes. Une remarque s'impose à ce propos: s'il est vrai que, lorsqu'il choisit une méthode de résolution, le chercheur n'inclut pas formellement son problème dans une classe quelconque, il ne s'en guide pas moins sur ses particularités telles que la forme de l'équation différentielle, la présence d'une singularité de la solution, le nombre de dérivées finies, etc. En choisissant l'algorithme, le chercheur tient compte de ces particularités et, qu'il le veuille ou non, situe le problème dans une classe possédant ces propriétés.

L'analyse des problèmes et leur classification appellent souvent la question suivante: à quelle classe donner la priorité? en particulier, quelle classe doit-on avoir en vue en rédigeant tel programme standard d'intégration numérique? Soit, par exemple, un programme qui calcule avec une fréquence sensiblement égale les intégrales de fonctions régulières et de celles qui ne le sont pas suffisamment (c'est une classification arbitraire, bien sûr). Quelles fonctions présideront au choix de l'algorithme? Un algorithme travaillant sur des fonctions régulières sera plus efficace, et le temps de résolution sera donc inférieur à celui du cas de fonctions ne jouissant pas de cette propriété. Un gain de temps de 50 % lors du calcul d'intégrales de fonctions régulières est moins avantageux que le même gain obtenu dans le cas des fonctions non régulières, ce qui fait qu'on prêtera surtout attention à l'efficacité de l'algorithme pour ces dernières. On agira autrement bien sûr si le pourcentage des fonctions non régulières est faible.

Supposons qu'on recherche une méthode optimale de résolution d'une classe de problèmes afin de l'exploiter en calcul automatique. Au bout d'un certain temps on trouve un procédé soit proche de l'optimum soit meilleur que les procédés existants, et on le perfec-

tionne. A quel instant faut-il arrêter ce travail de raffinement et passer à l'élaboration de programmes standard de résolution? Si l'on se hâte trop et qu'on implante l'algorithme à peine trouvé, il se peut que sa qualité médiocre nous oblige sous peu à en créer de nouveaux. Par contre, si l'on fait traîner les choses en longueur, la résolution par des programmes disponibles coûtera trop cher en temps.

On pourrait certes objecter que la mise au point de programmes usuels de qualité permettra dans la suite d'économiser beaucoup de temps machine. C'est un fait, encore que le temps machine soit de moins en moins déficitaire et onéreux. D'autre part, plus vite un problème est résolu, plus grand est le profit réel de la société dans son ensemble.

Ne perdons pas de vue que l'implantation rapide des algorithmes nouvellement créés est importante pour la recherche même des méthodes optimales de résolution. En effet, l'analyse des résultats numériques peut suggérer de remplacer la classe en question et ouvrir ainsi la voie à de nouvelles études théoriques. L'approche du problème d'optimalité doit donc être dynamique, à savoir on doit construire des méthodes optimales ou voisines de l'optimum pour des classes toujours nouvelles de problèmes posés par la science et la technique compte tenu des latitudes offertes par les machines à calculer électroniques.

Ceci étant, le travail courant est nécessaire au même titre que le dégrossissage de la solution. Science appliquée, le calcul mathématique se caractérise par la situation suivante: les problèmes de type nouveau sont en faible quantité, mais ils impliquent une solution immédiate à tout prix. A cette étape, toute méthode acceptable fait l'affaire. Puis, les problèmes affluent toujours plus nombreux, on arrive à poser le problème d'optimisation des méthodes d'une façon plus ou moins satisfaisante et on trouve une solution. Enfin, à l'étape suivante, les problèmes sont résolus par des systèmes de programmes standard correspondants. Leur étude n'est pas finie pour autant: on ne saurait élaborer de méthodes efficaces de résolution de problèmes nouveaux sans repenser les anciens et les soumettre à un examen théorique.

Le rapport entre le travail courant et la planification perspective est lui aussi un des facteurs majeurs de l'activité d'un centre scientifique quelconque. Celui-ci doit faire face à tout instant à des exigences d'importance vitale: rentabilité lorsqu'il s'agit d'un régime d'autofinancement, bilan en fin d'exercice, etc. Il est des moments critiques où l'on demande de résoudre à temps de nouvelles classes de problèmes. En planifiant, on préverra donc des recherches théoriques supplémentaires, des réserves de main-d'œuvre et de temps machine.

Revenons à l'optimisation des méthodes.

Le mathématicien qui a élaboré une méthode optimale ou voisine

de l'optimum se plaint souvent des difficultés qu'il éprouve pour la mettre en pratique. La cause en est souvent le conservatisme des praticiens, leur attachement à des techniques anciennes qui ont fait leurs preuves. Parfois, c'est la méthode même qui laisse à désirer. Il arrive par exemple qu'en plus (voire au lieu) de son optimalité, on exige d'une méthode qu'elle soit simple et permette de tester avec certitude les résultats obtenus. Il se peut également que la classe des problèmes étudiés ne coïncide pas avec celle de la plupart des problèmes posés réellement.

Autre chose encore. Si on sait qu'à l'heure actuelle et à l'avenir on n'aura à s'occuper que de peu des problèmes de la classe considérée, l'élaboration d'algorithmes optimaux et de programmes standard de résolution peut ne pas se justifier. D'autre part, l'étude du problème d'optimisation des méthodes sur des classes différentes est souvent utile parce que sa discussion fait apparaître de nouveaux procédés qui s'appliquent largement à la résolution de problèmes d'autres classes. Un exemple en est l'élaboration des méthodes numériques d'intégration.

§ 8. Formules de quadrature optimales sur des classes de fonctions à une dérivée

Les raisonnements du § 7 nous convainquent de l'importance que présente l'étude des différentes positions du problème d'optimisation des méthodes sur les classes de problèmes.

Soit à calculer l'intégrale

$$I(f) = \int_{\Omega} f(P)\, p(P)\, dP.$$

Le domaine d'intégration Ω et la fonction de poids $p(P)$ sont supposés fixes. On détermine la classe de problèmes considérés par la donnée de la classe F de fonctions à intégrer. On appelle *erreur* dans la formule de quadrature

$$I(f) \approx S_N(f) = \sum_{j=1}^{N} D_j f(P_j)$$

sur la classe F, la quantité

$$R_N(F) = \sup_{f \in F} |R_N(f)|,$$

où, comme d'habitude

$$R_N(f) = I(f) - S_N(f).$$

La borne inférieure

$$W_N(F) = \inf_{D_j, P_j} R_N(F)$$

est dite *estimation optimale de l'erreur dans les formules de quadrature sur la classe considérée.* S'il existe une formule de quadrature telle que $R_N(F) = W_N(F)$, on dit qu'elle est *optimale sur la classe considérée.*

Prenons

$$I(f) = \int_0^1 f(x)\,dx, \quad f \in F = C_1(A, [0, 1]),$$

$C_1(A, [0, 1])$ étant une classe de fonctions continues de dérivée première continue par morceaux vérifiant la condition $|f'(x)| \leqslant A$. La somme de quadrature s'écrit

$$S_N(f) = \sum_{j=1}^{N} D_j f(x_j), \quad x_1 < \ldots < x_N.$$

Imposons la condition

$$\sum_{j=1}^{N} D_j = 1. \tag{1}$$

Si (1) n'est pas satisfaite, on a, pour $f(x) = c = \text{const}$,

$$R_N(f) = \Big(1 - \sum_{j=1}^{N} D_j\Big) c \neq 0.$$

Toutes les fonctions $f(x) = \text{const}$ appartiennent à la classe donnée, donc,

$$R_N(F) \geqslant \sup_c \Big(\Big|1 - \sum_{j=1}^{N} D_j\Big|\,|c|\Big) = \infty.$$

Il est clair qu'on ne saurait parler de l'optimalité d'une formule pareille. La formule de quadrature vérifiant (1) est exacte pour des constantes, i.e. des polynômes de degré zéro. Selon (2.6) on a pour de telles formules

$$R_N(C_1(A, [0, 1])) = A \int_0^1 |K_N(y)|\,dy,$$

où

$$K_N(y) = 1 - y - \sum_{j=1}^{N} D_j \overline{(x_j - y)}^0,$$

$$\overline{(t)}^0 = \begin{cases} 1 & \text{pour } t > 0, \\ 0 & \text{pour } t \leqslant 0. \end{cases}$$

Si $x_{m-1} \leqslant y < x_m$, il vient

$$(\overline{x_j - y})^0 = \begin{cases} 0 \text{ pour } j < m, \\ 1 \text{ pour } j \geqslant m, \end{cases}$$

(nous avons posé $0^0 = 0$) ; aussi

$$\sum_{j=1}^{N} D_j (\overline{x_j - y})^0 =$$

$$= \begin{cases} 1 & \text{pour } y < x_1, \\ Q_m = \sum_{j=m}^{N} D_j & \text{pour } x_{m-1} \leqslant y < x_m, \ m = 2, \ \ldots, \ N, \\ 0 & \text{pour } x_N \leqslant y. \end{cases} \qquad (2)$$

Le problème initial de minimisation de l'erreur sur une classe se réduit donc à la résolution du problème suivant : approcher au mieux, dans la métrique L_1 : $\| f \| = \int_0^1 | f(x) | dx$, la fonction $1 - y$ par des fonctions de la forme (2). Désignons $\int_0^1 | K_N(y) | dy$ par $V(Q_2, \ldots, Q_N ; x_1, \ldots, x_N)$. On a

$$V(Q_2, \ldots, Q_N ; x_1, \ldots, x_N) =$$

$$= \int_0^{x_1} | y | dy + \sum_{m=2}^{N} \int_{x_{m-1}}^{x_m} | (1-y) - Q_m | dy + \int_{x_N}^{1} | 1 - y | dy. \qquad (3)$$

Un seul terme, à savoir

$$V(Q_m) = \int_{x_{m-1}}^{x_m} | (1-y) - Q_m | dy,$$

dépend de Q_m fixe. Une intégration immédiate donne

$$V(Q_m) = \begin{cases} \dfrac{(x_m - x_{m-1})^2}{2} + (Q_m - (1 - x_{m-1}))(x_m - x_{m-1}) \\ \qquad \text{pour } 1 - x_{m-1} \leqslant Q_m, \\ \dfrac{(1 - Q_m - x_m)^2 + (1 - Q_m - x_{m-1})^2}{2} \\ \qquad \text{pour } 1 - x_m \leqslant Q_m \leqslant 1 - x_{m-1}, \\ \dfrac{(x_m - x_{m-1})^2}{2} - (Q_m - (1 - x_m))(x_m - x_{m-1}) \\ \qquad \text{pour } Q_m \leqslant 1 - x_m. \end{cases}$$

D'où

$$V'(Q_m) > 0 \quad \text{pour } Q_m > 1 - \frac{x_{m-1}+x_m}{2},$$

$$V'(Q_m) < 0 \quad \text{pour } Q_m < 1 - \frac{x_{m-1}+x_m}{2},$$

par conséquent,

$$\min_{Q_m} V(Q_m) = V\left(1 - \frac{x_{m-1}+x_m}{2}\right) = \frac{(x_m - x_{m-1})^2}{4}. \qquad (4)$$

A part $V(Q_m)$ le second membre de (3) contient

$$\int_0^{x_1} |y|\,dy + \int_{x_N}^1 |1-y|\,dy = \frac{x_1^2 + (1-x_N)^2}{2}.$$

Il en résulte, compte tenu de (3) et (4), que

$$\inf_{Q_2, \ldots, Q_N} V(Q_2, \ldots, Q_N; x_1, \ldots, x_N) =$$

$$= \bar{V}(x_1, \ldots, x_N) = \frac{x_1^2 + (1-x_N)^2}{2} + \sum_{m=2}^{N} \frac{(x_m - x_{m-1})^2}{4}.$$

Ainsi,

$$\inf_{Q_2, \ldots, Q_N;\, x_1, \ldots, x_N} V(Q_2, \ldots, Q_N; x_1, \ldots, x_N) =$$

$$= \inf_{x_1, \ldots, x_N} \bar{V}(x_1, \ldots, x_N).$$

En identifiant à zéro les dérivées $\frac{\partial \bar{V}}{\partial x_m}$, on aboutit au système d'équations

$$\frac{3x_1 - x_2}{2} = 0, \quad \frac{3x_N - 2 - x_{N-1}}{2} = 0, \quad \frac{x_{m-1} - 2x_m + x_{m+1}}{2} = 0,$$

$$m = 2, \ldots, N-1,$$

dont la solution est

$$x_m = \frac{2m-1}{2N}, \quad m = 1, \ldots, N,$$

et, pour ces valeurs,

$$\bar{V}(x_1, \ldots, x_N) = \frac{1}{4N}.$$

Nous avons obtenu un point d'extrémum de la fonction $\bar{V}$, mais il ne correspond pas obligatoirement à sa borne inférieure; la fonction $\bar{V}(x_1, \ldots, x_N)$ est considérée dans le domaine $0 \leqslant x_1 < x_2 < \ldots < x_N \leqslant 1$, et si un certain point de la frontière réalise

sa borne inférieure, ce point ne satisfait pas forcément à la condition $\frac{\partial \overline{V}}{\partial x_m} = 0$.

Vérifions directement que la valeur trouvée de $\overline{V}$ est minimale. Portons $a_j \equiv 1$, $b_1 = x_1$, $b_2 = b_3 = \frac{x_2 - x_1}{2}$, $b_4 = b_5 = \frac{x_3 - x_2}{2}$, ..., $b_{2N-2} = b_{2N-1} = \frac{x_N - x_{N-1}}{2}$, $b_{2N} = 1 - x_N$ dans l'inégalité de Cauchy-Bouniakovski

$$\left|\sum_{j=1}^{2N} a_j b_j\right|^2 \leqslant \left(\sum_{j=1}^{2N} a_j^2\right)\left(\sum_{j=1}^{2N} b_j^2\right);$$

on a évidemment $\sum_{j=1}^{2N} a_j b_j = 1$. Finalement

$$1 \leqslant 2N \sum_{j=1}^{2N} b_j^2$$

ou

$$\frac{1}{4N} \leqslant \frac{1}{2}\left(\sum_{j=1}^{2N} b_j^2\right) = \overline{V}(x_1, \ldots, x_N).$$

Ainsi, la valeur minimale de $\int_0^1 |K_N(y)|\,dy$, qui est égale à $\frac{1}{4N}$, est atteinte pour

$$x_j = \frac{2j-1}{2N}, \quad Q_j = 1 - \frac{x_{j-1} + x_j}{2} = 1 - \frac{j-1}{N}.$$

On a pour la formule de quadrature correspondante

$$D_1 = 1 - Q_1 = \frac{1}{N}, \quad D_N = Q_N = \frac{1}{N},$$

$$D_j = Q_j - Q_{j+1} = \frac{1}{N}, \quad j = 2, \ldots, N-1.$$

On voit que la formule optimale sur la classe considérée est bien

$$\int_0^1 f(x)\,dx \approx \frac{1}{N} \sum_{j=1}^{N} f\left(\frac{2j-1}{2N}\right)$$

dont le reste est évalué par $\frac{A}{4N}$.

En qualité d'exemple prenons la classe $C_1^{(1)}(A, [0, 1])$ de fonctions continûment dérivables par morceaux et telles que $\int_0^1 |f'(x)|\,dx \leqslant A$.

Selon (6.10) l'estimation de l'erreur prend la forme

$$R_N(C_1^{(1)}(A, [0, 1])) \leqslant \sup_{[0, 1]} |K_N(y)| \int_0^1 |f'(x)|\,dx. \tag{5}$$

Montrons qu'il s'agit en fait d'une égalité. La fonction $K_N(y) = (1-y) - Q_m$ est linéaire sur $[x_{m-1}, x_m)$, aussi $\sup_{[x_{m-1}, x_m)} |K_N(y)| = \max(|1-x_{m-1}-Q_m|, |1-x_m-Q_m|)$, et $\sup_{[0, x_1)} |K_N(y)| = \sup_{[0, x_1)} |y| = x_1$, $\sup_{[x_N, 1]} |K_N(y)| = \sup_{[x_N, 1]} |1-y| = 1-x_N$.

Donc,

$$\sigma = \sup_{[0, 1]} |K_N(y)| = \max(x_1, |1-x_1-Q_2|, |1-x_2-Q_2|, \ldots, 1-x_N). \tag{6}$$

$K_N(y)$ étant linéaire sur chacun des intervalles $[0, x_1)$, $[x_1, x_2)$, ..., $[x_{N-1}, x_N)$, $[x_N, 1]$, on dit que quel que soit ε, on peut indiquer un segment $[y_1, y_2]$, $y_2 > y_1$, appartenant à l'un des segments mentionnés et tel que

$$|K_N(y)| \geqslant \sup |K_N(y)| - \varepsilon, \quad \operatorname{sg} K_n(y) = \text{const}.$$

Soit

$$f_\varepsilon(x) = \begin{cases} 0 & \text{pour } x \leqslant y_1, \\ A\dfrac{x-y_1}{y_2-y_1} & \text{pour } y_1 \leqslant x \leqslant y_2, \\ 1 & \text{pour } x \geqslant y_2. \end{cases}$$

Alors $f'_\varepsilon(x) = \dfrac{A}{y_2-y_1}$ pour $x \in (y_1, y_2)$ et $f'_\varepsilon(x) = 0$ pour $x \notin [y_1, y_2]$. L'erreur sur cette fonction est

$$R_N(f_\varepsilon) = \int_{y_1}^{y_2} \frac{K_N(y)\,A}{y_2-y_1}\,dy,$$

et, par conséquent,

$$|R_N(f_\varepsilon)| \geqslant A\left(\sup_{[0, 1]} |K_N(y)| - \varepsilon\right).$$

ε étant arbitraire, il vient

$$R_N(C_1^{(1)}(A, [0, 1])) \geqslant A \sup_{[0, 1]} |K_N(y)|.$$

Il en résulte, compte tenu de (5),

$$R_N(C_1^{(1)}(A, [0, 1])) = A \sup_{[0, 1]} |K_N(y)|.$$

Il est immédiat que

$$\{x_1\} + (\{1-x_1-Q_2\} + \{-(1-x_2-Q_2)\}) + \ldots$$
$$\ldots + (\{1-x_{N-1}-Q_N\} + \{-(1-x_N-Q_N)\}) + \{1-x_N\} =$$
$$= x_1 + (x_2-x_1) + \ldots + (x_N-x_{N-1}) + 1 - x_N = 1.$$

Selon (6), le maximum des modules des termes entre accolades est $\sup |K_N(y)|$. Il est au moins $\dfrac{1}{2N}$ pour une somme de $2N$ termes égale à 1. Aussi

$\sup_{[0,1]} |K_N(y)| \geqslant \frac{1}{2N}$. Le maximum des modules est $\frac{1}{2N}$ si et seulement si les termes sont tous égaux à $\frac{1}{2N}$. S'il existe donc une formule de quadrature pour laquelle

$$\sup_{[0,1]} |K_N(y)| = \frac{1}{2N},$$

on a obligatoirement

$$x_1 = \frac{1}{2N}, \quad 1 - x_1 - Q_2 = \frac{1}{2N}, \quad Q_2 + x_2 - 1 = \frac{1}{2N}, \ldots, \quad 1 - x_N = \frac{1}{2N}.$$

Portant $x_1 = \frac{1}{2N}$ dans la deuxième équation, il vient $Q_2 = 1 - \frac{1}{N}$; portant cette valeur de Q_2 dans la troisième, on obtient $x_2 = \frac{3}{2N}$, et ainsi de suite. Finalement

$$x_j = \frac{2j-1}{2N}, \quad Q_j = 1 - \frac{j-1}{N}.$$

Il correspond à ces Q_j les mêmes $D_j = \frac{1}{N}$. Ainsi, la formule

$$\int_0^1 f(x)\,dx \approx \frac{1}{N} \sum_{j=1}^{N} f\left(\frac{2j-1}{2N}\right)$$

avec un reste évalué par $\frac{A}{2N}$ est bien optimale sur la classe considérée.

A l'heure actuelle, les formules de quadrature optimales ne sont établies que pour un jeu restreint de classes de fonctions, principalement d'une variable. Leur intérêt pratique pour les applications est minime, ce qui ne signifie nullement que leur construction doit être négligée.

Les représentants d'autres disciplines, des sciences humaines en premier lieu, disent que la mathématique est une discipline quantitative, que les « mathématiciens calculent » tandis que leurs sciences à eux s'occupent de déductions qualitatives.

Rien n'est plus faux. Il est vrai que la mathématique influence de nombreux aspects de la vie sociale en les pourvoyant de résultats numériques de toutes sortes tels la sécurité de constructions, les plans optimaux, etc. Il est également indéniable qu'elle est l'outil par excellence de l'analyse qualitative de phénomènes se déroulant dans la nature et la société. Il n'est pas rare que ces phénomènes soient simulés par un grand nombre de modèles contradictoires qui ne se prêtent pas à un triage théorique. Seul le calcul du modèle sur ordinateur détermine son degré de conformité avec la réalité, et les conclusions des mathématiciens sont en l'occurrence qualitatives par

essence. La valeur de certains modèles mathématiques de l'économie se mesure souvent non pas au plan optimal qui résulte de leur examen, mais à une recommandation essentiellement qualitative portant disons sur l'orientation nouvelle à donner à la structure de la balance des combustibles, au système de paiement afin de stimuler la branche correspondante.

De même, on apprécie les constructions ci-dessus des formules de quadrature optimales et leur généralisation au cas de fonctions très régulières de plusieurs variables non pas d'après les formules concrètes en résultant, mais d'après les éclaircissements qu'elles apportent à l'aspect qualitatif du problème : où telle méthode est la plus performante, quelle précision escompter en utilisant telle information sur la fonction à intégrer, quelle est la densité de répartition des points dans les bonnes formules?

Nous avons vu, par exemple, que seule l'information sur le caractère borné de la dérivée première permettait une estimation d'erreur en $O\left(\frac{1}{N}\right)$ au plus. Ainsi, les évaluations obtenues au § 6 pour $n = 1$ ue peuvent être améliorées par rapport à N. Nous avons vu, d'autre part, que l'hypothèse de la dérivée première bornée ne pouvait pas améliorer de beaucoup l'estimation de l'erreur obtenue en supposant que

$$\int_0^1 |f'(x)|\,dx$$

était bornée.

§ 9. Optimisation de la répartition des points de base des formules de quadrature

Les méthodes d'intégration numérique auraient pu se développer par la création de méthodes optimales sur diverses classes $C_r^{(p)}(A, [a, b])$ et l'élaboration de programmes adéquats. On s'est vite aperçu que pour ces classes les procédés connus avec estimation (6.7), (6.12) étaient proches de l'optimum ; nous verrons au § 7, ch. V que l'optimisation de la formule de quadrature n'améliore pas l'ordre de l'estimation (6.7), (6.12) de l'erreur sur les classes considérées. On a également constaté que certaines méthodes qui coïncident pratiquement avec les méthodes classiques d'Euler et de Gregory sont *asymptotiquement optimales* relativement à l'évaluation de la partie principale de l'erreur. Plus précisément, dans ces méthodes sur les classes correspondantes de fonctions l'erreur a été évaluée par

$$\frac{c_r}{N^r} + \frac{c_{r+1}}{N^{r+1}},$$

tandis qu'elle l'est par

$$\frac{c_r}{N^r}+O\left(\frac{1}{N^{r+1}}\right)$$

pour des méthodes optimales sur ces classes.

On pourrait penser que du moment qu'il existe des méthodes presque optimales, l'étape suivante consisterait à les faire accepter par tous les systèmes d'intégration.

Cette affirmation est sujette à caution. En effet, l'extrême diversité des problèmes à résoudre implique une certaine prudence lors de la mise en pratique des méthodes. Nous ne sommes pas tout à fait certains que notre description de ces classes de problèmes correspond le mieux à une classe de problèmes réels ; ainsi, les classes C_1 régissent mal les problèmes réels. Devant l'attachement aux méthodes admises, les partisans de nouvelles techniques doivent faire preuve d'une énergie persévérante. Ils ne doivent pas oublier pour autant que les vieilles techniques ont fait leurs preuves et sont peut-être plus appropriées à des problèmes de certaines classes. Avant de les abandonner on procédera donc à une analyse pratique et théorique suffisante.

Dans notre cas (problème d'optimisation des méthodes d'intégration), avant même d'éclaircir complètement les questions afférentes à l'optimalité des méthodes, on est arrivé aux conclusions suivantes.

En divisant le segment d'intégration de départ en segments élémentaires égaux $a_q - a_{q-1} \equiv H$, nous obtenons l'information sur la fonction à intégrer de façon uniforme sur tout le segment. D'autre part, la fonction sur tel tronçon peut s'avérer plus régulière, ce qui nécessitera relativement moins de points de base. Aussi, dans les méthodes pratiquement optimales, le partage du segment d'intégration se conformera au comportement de la fonction à intégrer.

Voyons comment répartir les points de base d'après le comportement de la dérivée de la fonction à intégrer. Il existe plusieurs façons qui ont de nombreux traits communs, mais tel problème appelle telle approche. Pour simplifier l'exposé, bornons-nous au cas de la formule des trapèzes.

Soit à calculer

$$I(f)=\int_0^1 f(x)\,dx$$

avec $f(x)$ telle que $|f''(x)|\leqslant A_l$ sur (B_{l-1}, B_l), $l=1$, $0=B_0<B_1<\ldots<B_q=1$. On trouve

$$I_l(f)=\int_{B_{l-1}}^{B_l} f(x)\,dx$$

par la formule des trapèzes généralisée avec les segments partiels égaux de longueur $H_l = \frac{b_l}{N_l}$, $b_l = B_l - B_{l-1}$:

$$I_l(f) \approx S_l(f) = H_l \left(\frac{f(B_{l-1})}{2} + f(B_{l-1} + H_l) + \ldots + \frac{f(B_l)}{2} \right).$$

Conformément à (6.8), le reste est évalué par $\frac{A_l b_l^3}{12 N_l^2}$. L'erreur totale en remplaçant $I(f)$ par la somme $\sum_{l=1}^{q} S_l(f)$ ne dépassera pas

$$\Phi = \sum_{l=1}^{q} \frac{A_l b_l^3}{12 N_l^2}.$$

Nous nous proposons de choisir les quantités N_l de façon que, pour le nombre N donné ($N = N_1 + \ldots + N_q$) de segments de division, l'erreur Φ admette une estimation totale minimale.

Trouvons le minimum de Φ dans la condition $\Psi = N_1 + \ldots + N_q - N = 0$ sans supposer N_l entiers. En égalant à zéro les dérivées de la fonction de Lagrange $\Phi + \lambda\Psi$, on obtient le système d'équations

$$\frac{\partial(\Phi + \lambda\Psi)}{\partial N_l} = -\frac{A_l b_l^3}{6 N_l^3} + \lambda = 0.$$

Il s'ensuit

$$N_l = b_l \sqrt[3]{\frac{A_l}{6\lambda}}. \tag{1}$$

La condition $\Psi = 0$ fournit l'équation par rapport à λ:

$$\left(\sum_{l=1}^{q} b_l \sqrt[3]{\frac{A_l}{6\lambda}} \right) = N.$$

Déterminons λ, puis à l'aide de (1), trouvons N_l. Puisque N_l doivent être entiers, prenons

$$N_l = N_l^0 = \left[b_l \sqrt[3]{\frac{A_l}{6\lambda}} \right]. \tag{2}$$

On a visiblement

$$N - q \leqslant N_1^0 + \ldots + N_q^0 \leqslant N.$$

Nous n'avons pas obtenu de minimum réel de Φ sur l'ensemble des $N_1, \ldots, N_q$ entiers quelconques tels que ($N_1 + \ldots N_q = N$), mais nous estimons superflue toute amélioration ultérieure.

Voici un autre problème variationnel qui présente généralement un intérêt pratique. Trouver un minimum de $N = N_1 + \ldots N_q$

tel que l'évaluation de l'erreur Φ est ε au plus. Comme Φ diminue de façon monotone avec la croissance des N_l, le cas $\Phi = \varepsilon$ suffit amplement. Mettons la fonction de Lagrange sous la forme

$$N_1 + \ldots + N_q + \lambda^{-1}(\Phi - \varepsilon).$$

En annulant ses dérivées par rapport à N_l, on aboutit aux mêmes équations (1). Substituant $N_l = b_l \sqrt[3]{\frac{A_l}{6\lambda}}$ dans $\Phi = \varepsilon$, il vient

$$\lambda^{2/3} \sum_{l=1}^{q} \left(\frac{A_l}{48}\right)^{1/3} b_l = \varepsilon.$$

Déterminons λ, puis les N_l correspondants. Les deux problèmes nous ont donné les mêmes relations en N_l. Pour pouvoir nous prononcer qualitativement sur la répartition optimale des points il nous suffira donc d'examiner l'un d'eux.

Notre but est d'élaborer des méthodes optimales de résolution pour ensuite, sur leur base, mettre au point des systèmes de programmes de résolution des problèmes mathématiques types. Soit un programme de calcul d'une intégrale avec une précision donnée qui se déroule comme suit. On calcule le tableau des valeurs de la fonction sur un réseau $x_0, \ldots, x_n$. Moyennant ce tableau on dresse celui des différences divisées

$$f(x_0; x_1; x_2), \ldots, f(x_{n-2}; x_{n-1}; x_n).$$

Son étude détermine le partage le plus rationnel en segments partiels (B_{l-1}, B_l) et les valeurs correspondantes de A_l. Ensuite, compte tenu de (1), on choisit les N_l et on effectue l'intégration. Il existe des programmes en exploitation qui réalisent l'algorithme de calcul de l'intégrale ne différant que très peu de l'algorithme décrit. Il paraît que de pareils programmes sont très efficaces dans la résolution des problèmes de certaines classes.

Néanmoins la plupart des algorithmes des programmes standard en usage sur ordinateurs n'utilisent pas directement les relations obtenues, mais une conclusion qualitative, résultant de (1). Mettons cette égalité sous la forme

$$\frac{A_l}{12}\left(\frac{b_l}{N_l}\right)^3 = \frac{\lambda}{2}.$$

Son premier membre est l'évaluation de l'erreur sur un segment élémentaire de longueur $\frac{b_l}{N_l}$ de $[B_{l-1}, B_l]$. Ainsi, cette relation veut dire que pour *une répartition optimale des points d'intégration, les évaluations des erreurs sur les segments élémentaires coïncident.*

Pour arriver à ce résultat il suffisait de se borner au cas $q = 2$. Cette circonstance confirme la propriété suivante commune aux

caractéristiques qualitatives des méthodes de résolution des problèmes (pas forcément mathématiques): *pour les obtenir il suffit d'examiner les modèles les plus simples reflétant les aspects essentiels du phénomène.*

Voici une autre position semblable du problème relatif à l'optimisation de la répartition des points d'intégration. Afin de ne pas vous importuner par les détails, nous n'allons évaluer les termes d'ordre supérieur dans l'évaluation d'erreur que sous une forme sommaire.

Supposons que le segment d'intégration [0, 1] soit partagé en parties $[a_{q-1}, a_q]$, $q = 1, \ldots, N$, $a_0 = 0$, $a_N = 1$, et que l'intégrale pour chaque segment élémentaire est calculée par la formule des trapèzes

$$I_q(f) = \int_{a_{q-1}}^{a_q} f(x)\,dx \approx s_q(f) = \frac{a_q - a_{q-1}}{2}\,(f(a_{q-1}) + f(a_q)).$$

L'intégrale sur le segment [0, 1] tout entier est alors déterminée par

$$I(f) \approx \sum_{q=1}^{N} s_q(f)$$

et le reste s'évalue comme suit:

$$r = \sum_{q=1}^{N} \Big(\max_{[a_{q-1},\, a_q]} |f''(x)| \Big) \frac{(a_q - a_{q-1})^3}{12}. \tag{3}$$

Soit $|f''(x)| \leqslant F(x)$ sur [0, 1], $F(x)$ continue; prenons pour a_q les valeurs $\varphi\left(\frac{q}{N}\right)$ d'une fonction φ continûment dérivable et telle que $\varphi(0) = 0$, $\varphi(1) = 1$. Il est immédiat que

$$a_q - a_{q-1} = \varphi\left(\frac{q}{N}\right) - \varphi\left(\frac{q-1}{N}\right) = \varphi'\left(\frac{q}{N}\right)\frac{1}{N} + o\left(\frac{1}{N}\right) \quad \text{pour } N \to \infty,$$

aussi

$$\max_{[a_{q-1},\, a_q]} |f''(x)| \leqslant \max_{[a_{q-1},\, a_q]} F(x) = F(a_q) + o(1) = F\left(\varphi\left(\frac{q}{N}\right)\right) + o(1).$$

Ces relations entraînent

$$\frac{(a_q - a_{q-1})^3}{12} \Big(\max_{[a_{q-1},\, a_q]} |f''(x)| \Big) \leqslant \varepsilon_q =$$

$$= \left(\varphi'\left(\frac{q}{N}\right)\right)^3 \frac{F\left(\varphi\left(\frac{q}{N}\right)\right)}{12N^3} + o\left(\frac{1}{N^3}\right).$$

En portant ces dernières dans (3), il vient

$$r \leqslant \bar{r} = \frac{1}{N^2} \left\{ \sum_{q=1}^{N} \frac{1}{N} \left(\varphi'\left(\frac{q}{N}\right)\right)^3 \frac{F\left(\varphi\left(\frac{q}{N}\right)\right)}{12} \right\} + o\left(\frac{1}{N^2}\right).$$

L'expression entre accolades est la somme de quadratures de Riemann pour l'intégrale

$$\int_0^1 (\varphi'(t))^3 \frac{F(\varphi(t))}{12}\, dt$$

d'une fonction continue. Par conséquent,

$$\bar{r} = \frac{1}{N^2}\left(\int_0^1 (\varphi'(t))^3 \frac{F(\varphi(t))}{12}\, dt\right) + o\left(\frac{1}{N^2}\right). \tag{4}$$

Proposons-nous de minimiser le terme principal (le premier) de (4). L'équation d'Euler se résout plus facilement si l'on adopte φ pour variable indépendante. Le coefficient de $\frac{1}{N^2}$ de la partie principale de l'erreur s'écrit alors

$$\int_0^1 (t'(\varphi))^{-2} \frac{F(\varphi)}{12}\, d\varphi.$$

L'équation d'Euler pour la fonction qui rend minimale la fonctionnelle

$$\int_0^1 G(\varphi, t, t')\, d\varphi$$

est

$$\frac{d}{d\varphi}\left(\frac{\partial G}{\partial t'}\right) - \frac{\partial G}{\partial t} = 0. \tag{5}$$

Dans le cas considéré $\frac{\partial G}{\partial t} = 0$ et (5) donne lieu à $\frac{\partial G}{\partial t'} = \text{const}$. En portant une valeur concrète de la fonction G, on obtient

$$(t'(\varphi))^{-3} \frac{F(\varphi)}{6} = \text{const}$$

ou

$$F(\varphi)(\varphi'(t))^3 = C_1. \tag{6}$$

La solution générale de cette équation dépend de C_1 et d'une autre constante C_2. Les valeurs de C_1 et C_2 se déterminent à partir des conditions aux limites $\varphi(0) = 0$, $\varphi(1) = 1$. La mise en pratique de la solution de ce problème variationnel est des plus variées. Il existe par exemple des programmes réalisant l'intégration numérique (6) sur un réseau à pas sensiblement supérieur à $\frac{1}{N}$, puis répartissant les points selon la solution obtenue.

La relation (6) autorise la même conclusion sur l'égalité des évaluations des erreurs sur les segments élémentaires pour une répartition

optimale des points. En effet, multiplions (6) par $\frac{1}{12N^3}$, posons $t = \frac{q}{N}$ et remplaçons $F\left(\varphi\left(\frac{q}{N}\right)\right)$ et $\frac{1}{N}\varphi'\left(\frac{q}{N}\right)$ respectivement par les quantités équivalentes $\max\limits_{[a_{q-1},\, a_q]} F(x)$, $a_q - a_{q-1}$. Le résultat en est

$$\left(\max_{[a_{q-1},\, a_q]} F(x)\right) \frac{(a_q - a_{q-1})^3}{12} \approx \frac{C_1}{12N^3}. \tag{7}$$

Voici un autre moyen d'utiliser l'équation (6). Soit à calculer une grande série d'intégrales dont les fonctions sous le signe somme présentent une même allure caractéristique. Isolons une fonction modèle élémentaire pour laquelle le problème d'optimisation de points se résout de façon explicite et intégrons pour une répartition de points correspondant à cette fonction. Si la variation des fonctions de la série considérée dépend d'un certain paramètre, on en tiendra compte lors du choix de la fonction modèle qui peut ne pas appartenir à la classe intéressée. Plus on a de problèmes à résoudre et plus vite s'amortissent les dépenses nécessitées par un bon choix et une étude réussie du problème modèle.

§ 10. Exemples d'optimisation de la répartition des points d'intégration

Voyons comment se présente la solution de (9.6) dans certains cas concrets.

EXEMPLE 1. Soit à calculer la série d'intégrales

$$I(b) = \int_0^1 f(b, x)\, dx,$$

b est le paramètre de la série, $f(b, x) = x^b g(b, x)$, $-1 < b < 2$, $g(b, x)$ est régulière, $g(b, 0) \neq 0$. Si $b \neq 0$, 1, la dérivée seconde $f_{xx}(b, x)$ n'est pas bornée dans le voisinage du point 0; on tiendra compte de cela lors du choix du problème modèle. Dans le voisinage de $x = 0$, on a

$$f_{xx}(b, x) = b(b-1)\, x^{b-2} g(b, 0) + O(x^{b-1}).$$

Donc f_{xx} y est proportionnelle à la dérivée seconde de la fonction x^b, et par suite il est naturel de prendre cette dernière pour modèle. Adoptons en qualité de $F(x)$ la quantité $|b(b-1)|\, x^{b-2}$, l'équation (9.6) s'écrit alors

$$|b(b-1)|\, \varphi^{b-2} \left(\frac{d\varphi}{dt}\right)^3 = C_1.$$

D'où

$$\left(\frac{3\sqrt[3]{|b(b-1)|}}{b+1}\right)\varphi^{\frac{b+1}{3}} = C_1 t + C_2;$$

la condition $\varphi(0)=0$ donne $C_2=0$, la condition $\varphi(1)=1$ entraîne $\varphi^{\frac{b+1}{3}} = t$. Ainsi

$$\varphi(t) = t^{\frac{3}{1+b}} \tag{1}$$

et le problème modèle de calcul de $\int_0^1 x^b dx$ admet $a_q = \left(\frac{q}{N}\right)^{\frac{3}{1+b}}$ pour répartition de points optimale dans le sens considéré.

Nos raisonnements ci-dessus manquent de rigueur, car l'évaluation (9.4) suppose que l'évaluation de la dérivée seconde de $F(x)$ est bornée, or cette condition n'est pas réalisée dans le problème modèle étudié.

Essayons de démontrer que dans l'exemple considéré il serait bon de se guider sur (9.6) pour choisir la répartition optimale des points de $\varphi(t)$. Soit $F(x)$ continue en tous les points de (0, 1] (hypothèse de l'exemple 1). Supposons que la répartition correspondant à une certaine fonction $\varphi_0(t)$ est telle que la partie principale de l'estimation de l'erreur dans l'intégration existe, ne dépasse pas celle de n'importe quelle autre répartition et est de l'ordre de $O(1/N^2)$. Montrons que $\varphi_0(t)$ est alors solution de (9.6). A cette fin prenons un η quelconque supérieur à 0. La partie principale de l'erreur d'intégration sur $[\varphi_0(\eta), 1]$ a la forme $\frac{I_\eta(\varphi_0)}{12N^2}$, avec

$$I_\eta(\varphi) = \int_\eta^1 F(\varphi(t))(\varphi'(t))^3 dt.$$

Soit $\varphi(t)$ une fonction arbitraire de la classe considérée qui se confond avec $\varphi_0(t)$ si $t \leqslant \eta$. Il est clair que les erreurs d'intégration sur $[0, \varphi(\eta)]$ sont les mêmes pour de telles fonctions. La partie principale correspondant à $\varphi_0(t)$ étant minimale, on a $I_\eta(\varphi) \geqslant I_\eta(\varphi_0)$. Il en résulte que $\varphi_0(t)$ vérifie l'équation (9.6) pour $\eta < t \leqslant 1$. η étant arbitraire, cette fonction la vérifie pour $0 < t \leqslant 1$.

Montrons sur l'exemple de l'intégrale $\int_0^1 x^b\,dx$ pour $-1 < b < 2$ que si $\varphi_0(t)$ est solution de (9.6) avec les conditions aux limites $\varphi_0(0) = 0$, $\varphi_0(1) = 1$, la partie principale de l'estimation d'erreur existe et est de l'ordre de $\frac{1}{N^2}$. Selon (1) la fonction correspondante $\varphi_0(t)$ est $t^{\frac{3}{1+b}}$; il est clair que $d = \frac{3}{1+b} > 1$. Conformément à la formule de Taylor

$$a_q - a_{q-1} = \varphi_0'\left(\frac{q}{N}\right)\frac{1}{N} - \varphi_0''(\theta_q)\frac{1}{2N^2}, \tag{2}$$

avec $\frac{q-1}{N} \leqslant \theta_q \leqslant \frac{q}{N}$. Puisque $\varphi_0''(t) = d(d-1)t^{d-2} > 0$, on a

$$a_q - a_{q-1} \leqslant \varphi_0'\left(\frac{q}{N}\right)\frac{1}{N} = d\left(\frac{q}{N}\right)^{d-1}\cdot\frac{1}{N}. \tag{3}$$

Pour $2 \leqslant q$ il vient

$$|\varphi_0''(\theta_q)| \leqslant \max\left(\left|\varphi_0''\left(\frac{q-1}{N}\right)\right|, \left|\varphi_0''\left(\frac{q}{N}\right)\right|\right) = O\left(\left(\frac{q}{N}\right)^{d-2}\right). \tag{4}$$

Il découle de (2) et (4)

$$a_q - a_{q-1} = d\left(\frac{q}{N}\right)^{d-1}\frac{1}{N} + O\left(\left(\frac{q}{N}\right)^{d-2}\frac{1}{N^2}\right) =$$

$$= \left(d\left(\frac{q}{N}\right)^{d-1}\frac{1}{N}\right)\left(1 + O\left(\frac{1}{q}\right)\right) \quad \text{si} \quad 2 \leqslant q.$$

Dans le cas considéré,

$$F(x) = |b(b-1)|x^{b-2}, \quad |F'(x)| = |b(b-1)(b-2)|x^{b-3}$$

croît, donc

$$\max_{[a_{q-1}, a_q]}\left|F(x) - |b(b-1)|\left(\frac{q}{N}\right)^{d(b-2)}\right| \leqslant |F'(a_{q-1})|(a_q - a_{q-1}).$$

Si $2 \leqslant q$, compte tenu de (3), on a

$$|F'(a_{q-1})(a_q - a_{q-1})| \leqslant \text{const}\left(\frac{q}{N}\right)^{d-1+d(b-3)}\frac{1}{N}.$$

D'où

$$\max_{[a_{q-1}, a_q]} F(x) = |b(b-1)|\left(\frac{q}{N}\right)^{d(b-2)} + O\left(\left(\frac{q}{N}\right)^{d(b-2)-1}\frac{1}{N}\right) =$$

$$= |b(b-1)|\left(\frac{q}{N}\right)^{d(b-2)}\left(1 + O\left(\frac{1}{q}\right)\right).$$

Par conséquent,

$$\max_{[a_{q-1}, a_q]} F(x)\frac{(a_q - a_{q-1})^3}{12} =$$

$$= \frac{|b(b-1)|}{12}\left(\frac{q}{N}\right)^{d(b-2)}\left(d\left(\frac{q}{N}\right)^{d-1}\frac{1}{N}\right)^3\left(1 + O\left(\frac{1}{q}\right)\right); \tag{5}$$

du moment que $b = \frac{3-d}{d}$, l'exposant de la puissance à laquelle on élève $\frac{q}{N}$ est

$$3(d-1) + d(b-2) = 3(d-1) + d\left(\frac{3(1-d)}{d}\right) = 0.$$

De cette façon

$$\max_{[a_{q-1}, a_q]} F(x)\frac{(a_q - a_{q-1})^3}{12} = \frac{d^3|b(b-1)|}{12N^3}\left(1 + O\left(\frac{1}{q}\right)\right) \quad \text{pour} \quad 2 \leqslant q.$$

Pour q grands, le deuxième membre dépend faiblement de q; c'est naturel car la fonction $\varphi(t)$ vérifiant (9.6) satisfait également à (9.7).

L'estimation de l'erreur d'intégration sur le segment $[0, a_1]$ qui s'écrit

$$(\max_{[0, a_1]} F(x)) \frac{(a_1-0)^3}{12}$$

n'est pas valable vu que $F(x)$ n'est pas bornée. Pour $b < 0$, la fonction à intégrer x^b devient infinie si $x = 0$. Supposons-la égale à zéro dans cette dernière condition. Alors, quel que soit b, la différence entre l'intégrale $\int_0^{a_1} x^b\,dx$ et la somme $s_1(f)$ s'écrit

$$\frac{a_1^{b+1}}{b+1} - \frac{a_1 \cdot a_1^b}{2} = \left(\frac{1}{b+1} - \frac{1}{2}\right) a_1^{b+1} = \left(\frac{1}{b+1} - \frac{1}{2}\right) \frac{1}{N^3}.$$

La somme de cette expression et des estimations (5) de l'erreur sur les autres segments est égale à

$$r = \frac{d^3 b\,|\,b-1\,|}{12N^2} + O\left(\sum_{q=1}^{N} \frac{1}{qN^3}\right) = \frac{d^3 b\,|\,b-1\,|}{12N^2} + O\left(\frac{\ln N}{N^3}\right). \qquad (6)$$

Nous avons obtenu la relation requise.

PROBLÈME. Soit

$$f(x) = \begin{cases} 0 & \text{pour } x = 0, \\ x^b & \text{pour } x \in (0, 1], \end{cases}$$

$-1 < b < 1$, et l'intégrale $\int_0^1 f(x)\,dx$ se calcule par la formule des trapèzes à pas constant $a_q - a_{q-1} \equiv \frac{1}{N}$. Démontrer que l'erreur totale vérifie

$$R_N(f) \sim \frac{D(b)}{N^{1+b}},$$

où $D(b) \neq 0$.

La comparaison de la dernière estimation avec (6) entraîne que s'agissant de fonctions dont les dérivées présentent des singularités, l'optimisation de la répartition des points d'intégration peut augmenter la vitesse de convergence.

EXEMPLE 2. On calcule la série d'intégrales

$$\int_0^1 x^b \ln x g(b, x)\,dx,$$

$g(b, x)$ étant régulière, $g(b, 0) \neq 0$, b le paramètre de la série. Puisque $\ln x$ a une singularité en 0, il paraît naturel de prendre $x^b \ln x$ pour fonction modèle. Sa dérivée seconde s'écrit

$$x^{b-2}(b(b-1)\ln x + (2b-1)).$$

Si $F(x) = |(x^b \ln x)''|$, l'équation (9.6) ne se résout pas par quadratures; essayons donc de la simplifier. Pour x tendant vers zéro,

$|\ln x|$ croît plus lentement que toute puissance $x^{-\varepsilon}$, $\varepsilon > 0$. Aussi, pour simplifier la solution de (9.6), posons $F(x) = \text{const} \cdot x^{b-2}$.

EXEMPLE 3. Soit à calculer la série d'intégrales

$$\int_0^1 \exp\left(-\frac{x}{b}\right) g(b, x)\, dx, \tag{7}$$

$g(b, x)$ est régulière, $g(b, 0) \neq 0$, b le paramètre de la série, susceptible de valeurs très petites. Pour b petits, la fonction à intégrer et ses dérivées varient brusquement au voisinage du point $x = 0$ à cause du facteur $\exp\left(-\frac{x}{b}\right)$. On optimisera donc la répartition des points d'intégration sur le problème modèle du calcul de $\int_0^1 \exp\left(-\frac{x}{b}\right) dx$. Faisons $F(x) = \left|\left(\exp\left(-\frac{x}{b}\right)\right)''\right|$. L'équation (9.6) se met sous la forme

$$\frac{1}{b^2} \exp\left(-\frac{\varphi}{b}\right) \left(\frac{d\varphi}{dt}\right)^3 = C_1.$$

D'où

$$1 - \exp\left(-\frac{\varphi}{3b}\right) = C_3 t + C_4.$$

La condition $\varphi(0) = 0$ entraîne $C_4 = 0$ et la condition $\varphi(1) = 1$ donne

$$1 - \exp\left(-\frac{\varphi}{3b}\right) = \left(1 - \exp\left(-\frac{1}{3b}\right)\right) t.$$

Il en résulte

$$\varphi(t) = 3b \ln \frac{1}{1 - \left(1 - \exp\left(-\frac{1}{3b}\right)\right) t}.$$

EXEMPLE 4. La série d'intégrales est celle de l'exemple précédent, mais $g(b, x) = xh(b, x)$, $h(b, 0) \neq 0$. On peut prendre $\int_0^1 x \exp\left(-\frac{x}{b}\right) dx$ pour l'intégrale modèle. Pour

$$F(x) = \left|\left(x \exp\left(-\frac{x}{b}\right)\right)''\right| = \left|\frac{x}{b^2} - \frac{2}{b}\right| \exp\left(-\frac{x}{b}\right)$$

l'équation (9.6) correspondante ne se calcule pas par des quadratures élémentaires. Si le problème est urgent, on se borne à la répartition optimale de points construite pour l'exemple 3.

EXEMPLE 5. On demande de calculer la série d'intégrales

$$\int_0^1 \exp\left(-\frac{x^2}{b^2}\right) g(b, x)\, dx,$$

$g(b, x)$ étant régulière, $g(b, 0) \neq 0$, b, paramètre de la série, pouvant être petit. Dans cette dernière hypothèse, la fonction à intégrer et ses dérivées varient brusquement au voisinage de $x = 0$ du fait de leur multiplication par $\exp\left(-\frac{x^2}{b^2}\right)$. Adoptons donc en qualité de problème modèle le calcul de l'intégrale $\int_0^1 \exp\left(-\frac{x^2}{b^2}\right) dx$. Soit $F(x) = \left|\left(\exp\left(-\frac{x^2}{b^2}\right)\right)'\right|$, alors au lieu de (9.6) il vient

$$\left|1-2\frac{\varphi^2}{b^2}\right| \exp\left(-\frac{\varphi^2}{b^2}\right)\left(\frac{d\varphi}{dt}\right)^3 = \frac{C_1 b^2}{2}.$$

D'où

$$\int \sqrt[3]{\left|1-2\frac{\varphi^2}{b^2}\right|} \exp\left(-\frac{\varphi^2}{3b^2}\right) d\varphi = \int \frac{\sqrt[3]{C_1 b^2}}{\sqrt[3]{2}} dt.$$

Cette intégrale ne s'explicitant pas, des simplifications s'imposent ; on remplace, par exemple, $\sqrt[3]{\left|1-2\frac{\varphi^2}{b^2}\right|}$ par 1. Avec $\frac{\varphi}{b}$ importants et l'erreur de substitution grande, la répercussion de celle-ci n'est pas tellement décisive vu la petitesse du facteur $\exp\left(-\frac{\varphi}{3b^2}\right)$. Cette simplification faite, $\varphi(t)$ s'exprimera par la fonction inverse de

$$f \int_0^u \exp(-v^2)\, dv.$$

Les problèmes analogues aux exemples 3, 4, 5 sont assez fréquents. Ainsi, en calculant les diagrammes de directivité des antennes on recherche les séries d'intégrales

$$\int_0^1 \exp(ibg(b, x))\, h(b, x)\, dx$$

pour b variant dans de vastes limites et les fonctions $g(b, x)$, $h(b, x)$ suffisamment régulières. Si b n'est pas très important, on trouve ces intégrales à l'aide des formules de quadrature connues. Les dérivées de la fonction à intégrer augmentent avec b, ce qui impose un nombre toujours plus grand de points d'intégration. Si b est très grand, on fait recours à la méthode du col ou à d'autres procédés asymptotiques. Les deux techniques s'avèrent cependant peu efficaces pour des valeurs « intermédiaires » du paramètre b : la première est trop lourde, la seconde peu précise. On agit donc quelquefois de la manière suivante : on transforme le contour d'intégration de façon qu'il coïncide avec les lignes de la plus grande pente $\exp(ibg(b, x))$, comme on le fait dans la méthode du col. Les intégrales des fonctions variant brusquement que l'on obtient sont analogues à celles des exemples 2, 3, 4. On optimise ensuite la répartition des points d'intégration et l'on se sert des formules de quadrature correspondantes.

On voit sur les exemples 3, 4 qu'une optimisation de la répartition de points à partir de (9.6) exige un chercheur assez qualifié. Dans le paragraphe 17 du présent chapitre c'est donc la machine qui s'en chargera.

§ 11. Partie principale de l'erreur

Plus haut (§§ 1, 2, 6, 8), nous avons examiné plusieurs formules de quadrature et évalué leur erreur en fonction du maximum ou des intégrales des dérivées. Ce faisant, nous n'avons pas supposé l'existence des dérivées d'ordre supérieur. Si la fonction à intégrer possède de telles dérivées, cela pourrait contrebalancer certains termes de l'erreur. Pour pouvoir discuter la qualité des évaluations et la valeur réelle de l'erreur, il faut isoler dans celle-ci la partie principale.

Revenons à la formule (1.3)

$$I(f)=\int_a^b f(x)\,p(x)\,dx \approx S_n(f)=\frac{b-a}{2}\sum_{j=1}^{n} D_j f\left(\frac{b+a}{2}+\frac{b-a}{2}d_j\right) \quad (1)$$

avec pour reste (1.1):

$$R_n(f)=I(f)-S_n(f)=\int_a^b p(x)\,(f(x)-L_n(x))\,dx.$$

Afin de se fixer les idées, nous nous bornerons au cas $p(x)=1$. Après la substitution $x=\frac{b+a}{2}+\frac{b-a}{2}t$, le reste s'écrit comme suit:

$$R_n(f)=\frac{b-a}{2}\int_{-1}^{1}\left(f\left(\frac{b+a}{2}+\frac{b-a}{2}t\right)-Q_n\left(f\left(\frac{b+a}{2}+\frac{b-a}{2}t\right)\right)\right)dt. \quad (2)$$

Nous avons introduit la notation

$$Q_n(g(t))=\sum_{j=1}^{n} g(d_j)\prod_{i\neq j}\frac{t-d_i}{d_j-d_i}.$$

Soit un m tel que la formule (1) soit stricte pour tous les polynômes de degré $m-1$ et non stricte pour les polynômes de degré m. Le procédé de construction de la formule d'interpolation (1.3) entraîne $m \geqslant n$. Substituons dans (2) le développement taylorien

$$f\left(\frac{b+a}{2}+\frac{b-a}{2}t\right)=$$
$$=\sum_{j=0}^{l-1} f^{(j)}\left(\frac{a+b}{2}\right)\frac{\left(\frac{b-a}{2}t\right)^j}{j!}+f^{(l)}(\zeta(t))\frac{\left(\frac{b-a}{2}t\right)^l}{l!} \quad (3)$$

pour un $l>m$ ($a\leqslant\zeta(t)\leqslant b$ si $-1\leqslant t\leqslant 1$). En groupant les termes contenant les mêmes $f^{(j)}$ en facteur, on a

$$R_n(f)=\sum_{j=0}^{l-1} f^{(j)}\left(\frac{b+a}{2}\right)\left(\frac{b-a}{2}\right)^{j+1}\gamma_j+\rho_l,$$

où

$$\gamma_j = \int_{-1}^{1} \frac{(t^j - Q_n(t^j))}{j!}\, dt,$$

$$\rho_l = \left(\frac{b-a}{2}\right)^{l+1} \int_{-1}^{1} \frac{f^{(l)}(\zeta(t))\, t^l - Q_n(f^{(l)}(\zeta(t))\, t^l)}{l!}\, dt.$$

Compte tenu de l'expression explicite de l'opérateur d'interpolation Q_n on trouve l'estimation grossière du dernier terme

$$|\rho_l| \leqslant \gamma_l^0 \max_{[a,\, b]} |f^{(l)}(x)| \left(\frac{b-a}{2}\right)^{l+1}$$

avec

$$\gamma_l^0 = \int_{-1}^{1} \left(|t|^l + \sum_{j=1}^{n} |d_j|^l \prod_{i \neq j} \left|\frac{t-d_j}{d_j-d_i}\right|\right) \Big/ l!\, dt.$$

Le coefficient de $f^{(j)}\left(\frac{b+a}{2}\right)$ est égal à l'erreur commise en calculant l'intégrale de $\frac{\left(x-\frac{b+a}{2}\right)^j}{j!}$ d'après la formule considérée. C'est pourquoi $\gamma_0 = \ldots = \gamma_{m-1} = 0$. Par ailleurs, $\gamma_m \neq 0$, sinon la formule de quadrature serait précise pour n'importe quel polynôme de degré m. Ces raisonnements nous autorisent à écrire

$$R_n(f) = \sum_{j=m}^{l-1} f^{(j)}\left(\frac{b+a}{2}\right)\left(\frac{b-a}{2}\right)^{j+1} \gamma_j + \\ + \max_{[a,\, b]} |f^{(l)}(x)| \left(\frac{b-a}{2}\right)^{l+1} \theta_l \gamma_l^0, \qquad (4)$$

où $|\theta_l| \leqslant 1$, $\gamma_m \neq 0$, $m \leqslant l-1$.

Rappelons-nous qu'au § 1, nous avons obtenu des relations de la forme

$$R_n(f) = \text{const} \cdot f^{(m)}(\zeta) \left(\frac{b-a}{2}\right)^{m+1} \qquad (5)$$

pour les formules des rectangles, des trapèzes et de Simpson. Pour la formule de Gauss avec $p(x) \equiv 1$, une telle relation découle de (3.9). Il résulte de (5) la validité d'une relation analogue à (4) avec $l = m + 1$.

Supposons que pour calculer

$$I(f) = \int_A^B f(x)\, dx$$

on applique la formule de quadrature (6.4)

$$I(f) \approx S_M^n(f) = \sum_{q=1}^{M} s_q(f),$$

où

$$s_q = \left(\frac{a_q - a_{q-1}}{2}\right) \sum_{j=1}^{n} D_j f\left(\frac{a_{q-1}+a_q}{2} + \frac{a_q - a_{q-1}}{a} d_j\right),$$

$a_q = \varphi\left(\frac{q}{M}\right)$, $\varphi(t)$ étant une fonction régulière, $\varphi(0) = A$, $\varphi(1) = B$. Alors $a_q - a_{q-1} = O\left(\frac{1}{M}\right)$. Utilisant (4) on écrit

$$\int_{a_{q-1}}^{a_q} f(x)\,dx - s_q = \gamma_m f^{(m)}\left(\frac{a_{q-1}+a_q}{2}\right)\left(\frac{a_q - a_{q-1}}{2}\right)^{m+1} + O\left(\frac{1}{M^{m+2}}\right).$$

Par conséquent, l'erreur totale sur tous les segments élémentaires est

$$R_M^n(f) = \gamma_m \sum_{q=1}^{M} f^{(m)}\left(\frac{a_{q-1}+a_q}{2}\right)\left(\frac{a_q - a_{q-1}}{2}\right)^{m+1} + O\left(\frac{1}{M^{m+1}}\right).$$

Effectuons des transformations analogues à celles du § 9 ;

$$a_q - a_{q-1} = \frac{1}{M}\varphi'\left(\frac{q}{M}\right) + O\left(\frac{1}{M^2}\right),$$

$$f^{(m)}\left(\frac{a_{q-1}+a_q}{2}\right) = f^{(m)}\left(\varphi\left(\frac{q}{M}\right)\right) + O\left(\frac{1}{M}\right).$$

On a, en définitive,

$$R_M^n(f) = \frac{(\gamma_m 2^{-(m+1)})}{M^m}\left\{\sum_{q=1}^{M} \frac{1}{M}\Phi\left(\frac{q}{M}\right)\right\} + O\left(\frac{1}{M^{m+1}}\right),$$

où $\Phi(t) = f^{(m)}(\varphi(t))\,(\varphi'(t))^{m+1}$. L'expression entre accolades représente la somme de quadratures de l'intégrale

$$\int_0^1 \Phi(t)\,dt,$$

qui est sa limite. Si $f^{(m+1)}(x)$ et $\varphi''(t)$ sont bornées, on montre que

$$R_M^n(f) = \frac{\gamma_m 2^{-(m+1)}}{M^m} \int_0^1 \Phi(t)\,dt + O\left(\frac{1}{M^{m+1}}\right). \qquad (6)$$

En particulier si les points $d_1, \ldots, d_n$ sont symétriques par rapport à 0 et que les dérivées $f^{(m-2)}(x)$ et $\varphi^{(3)}(t)$ soient bornées,

le deuxième terme à droite dans (6) est $O\left(\frac{1}{M^{m+2}}\right)$. Démontrons la dernière affirmation pour le cas de la formule des trapèzes à pas d'intégration constant, i.e. pour $\varphi_i' = \text{const}$ et $a_q - a_{q-1} \equiv H$. Soit $l = 4$; alors

$$\gamma_2 = \int_{-1}^{1} \frac{(t^2-1)}{2!}\, dt = -\frac{2}{3},$$

$$\gamma_3 = \int_{-1}^{1} \frac{(t^3-t)}{3!}\, dt = 0,$$

$$\gamma_4^0 = \int_{-1}^{1} \frac{\left(t^4 + \left|\frac{t-1}{2}\right| + \left|\frac{t+1}{2}\right|\right)}{4!}\, dt = \frac{1}{10}.$$

Ainsi,

$$\int_{a_{q-1}}^{a_q} f(x)\, dx - \frac{H}{2}\left(f(a_{q-1}) + f(a_q)\right) =$$

$$= -\frac{H^3}{12} f''\left(\frac{a_{q-1}+a_q}{2}\right) + \theta_q \frac{H^5}{320} \max_{[a_{q-1},\, a_q]} |f^{(4)}(x)|. \qquad (7)$$

L'expression $R_M^n(f)$ de l'erreur totale que l'on obtient en sommant (7) s'écrit

$$R_M(f) = -\frac{H^2}{12} E_2(f) + O\left((B-A)H^4\right),$$

où

$$E_2(f) = \sum_{q=1}^{M} Hf''\left(\frac{a_{q-1}+a_q}{2}\right).$$

La quantité $E_2(f)$ est la somme de quadratures pour l'intégrale $\int_A^B f''(x)\, dx$ si le calcul d'intégrales sur les segments élémentaires se fait à l'aide de la formule des rectangles (1.6):

$$\int_a^b g(x)\, dx \approx (b-a)\, g\left(\frac{b+a}{2}\right)$$

dont le reste s'évalue par

$$\left(\max_{[a,\, b]} |g''(x)|\right) \frac{(b-a)^3}{24}.$$

Par conséquent, en vertu de (6.6), on a

$$\left|E_2(f)-\int_A^B f''(x)\,dx\right|\leqslant\left(\max_{[A,\,B]}|f^{(4)}(x)|\right)\frac{(B-A)H^2}{24},$$

et, en définitive,

$$R_M(f)=-\frac{H^2}{12}\left(\int_A^B f''(x)\,dx\right)+O((B-A)H^4). \tag{8}$$

§ 12. Formules d'Euler et de Gregory

La relation (11.8) sert à plusieurs usages. Elle s'écrit, par exemple,

$$R_M(f)=-\frac{H^2}{12}(f'(B)-f'(A))+O((B-A)H^4). \tag{1}$$

Le calcul de $f'(B)-f'(A)$ fournit la partie principale de l'erreur. Supposons médiocre la précision obtenue. Mettons (1) sous la forme

$$I(f)=S_M^4(f)+O((B-A)H^4),$$

où

$$S_M^4(f)=S_M^2(f)-\frac{H^2}{12}(f'(B)-f'(A)).$$

L'expression de $S_M^4(f)$ est une formule de quadrature entachée de l'erreur $O((B-A)H^4)$, i.e. elle approche mieux $I(f)$ que la formule des trapèzes dont nous sommes partis.

Essayons d'isoler la partie principale de l'erreur dans la formule trouvée. On a

$$S_M^4(f)=\sum_{q=1}^{M}s_q^2(f),$$

$$s_q^2(f)=\frac{H}{2}(f(a_{q-1})+f(a_q))-\frac{H^2}{12}(f'(a_q)-f'(a_{q-1})).$$

$s_q^2(f)$ sera pris pour valeur approchée de l'intégrale

$$I_q(f)=\int_{a_{q-1}}^{a_q}f(x)\,dx.$$

Portons dans la différence $I_q(f)-s_q^2(f)$ la représentation de $f(x)$ sous forme d'une série partielle de Taylor (11.3) pour $l=6$. Comme au paragraphe 11, groupons les coefficients de $f^{(j)}\left(\frac{a_{q-1}+a_q}{2}\right)$:

$$I_q(f)-s_q^2(f)=\frac{H^5}{720}f^{(4)}\left(\frac{a_{q-1}+a_q}{2}\right)+O(H^7).$$

En sommant sur q il vient

$$I(f)-S_M^4(f)=\frac{H^4}{720}\left\{\sum_{q=1}^{N} Hf^{(4)}\left(\frac{a_{q-1}+a_q}{2}\right)\right\}+O((B-A)H^6).$$

L'expression entre accolades est la formule de quadrature des rectangles pour l'intégrale

$$\int_A^B f^{(4)}(x)\,dx=f^{(3)}(B)-f^{(3)}(A).$$

On a donc

$$I(f)-S_M^4(f)=\frac{H^4}{720}(f^{(3)}(B)-f^{(3)}(A))+O((B-A)H^6).$$

Nous avons obtenu le reste de la formule donnant $S_M^4(f)$ et une nouvelle formule de quadrature

$$I(f)\approx S_M^6(f)=S_M^2(f)-\frac{H^2}{12}(f'(B)-f'(A))+\frac{H^4}{720}(f^{(3)}(B)-f^{(3)}(A))$$

avec pour reste $O((B-A)H^6)$, et ainsi de suite. En continuant, on trouve, pour chaque p entier, la *formule de quadrature d'Euler*

$$I(f)\approx S_M^{2p}(f)$$

avec

$$S_M^{2p}(f)=S_M^2(f)-\sum_{k=1}^{p-1}\gamma_{2k}H^{2k}(f^{(2k-1)}(B)-f^{(2k-1)}(A)). \qquad (2)$$

Les nombres γ_j vérifient

$$\frac{x}{\exp x-1}=\sum_{j=0}^{\infty}\gamma_j x^j.$$

γ_j se notent d'ordinaire $B_j/j!$; B_j sont dits *nombres de Bernoulli*.

Afin d'évaluer le reste de (2) on procède comme suit. Soit

$$I_q(f)\approx s_q^p(f)=\frac{H}{2}(f(a_{q-1})+f(a_q))-\sum_{k=1}^{p-1}\gamma_{2k}H^{2k}(f^{(2k-1)}(a_q)-f^{(2k-1)}(a_{q-1})).$$

Portons dans le reste

$$r_q^p(f)=I_q(f)-s_q^p(f)$$

le développement de $f(x)$ sous forme d'une série taylorienne partielle

$$\sum_{j=0}^{2p-1} f^{(j)}\left(\frac{a_{q-1}+a_q}{2}\right)\frac{1}{j!}\left(x-\frac{a_{q-1}+a_q}{2}\right)^j+\int_{(a_{q-1}+a_q)/2}^{x} f^{(2p)}(y)\frac{(x-y)^{2p-1}}{(2p-1)!}\,dy.$$

Le reste s'écrit alors

$$r_q^p(f)=\int_{a_{q-1}}^{a_q} K_p\left(x-\frac{a_{q-1}+a_q}{2},\ H\right) f^{(2p)}(x)\,dx$$

(comme au § 2) ; les coefficients des dérivées d'ordre inférieur $f^{(j)}\left(\frac{a_{q-1}+a_q}{2}\right)$, $j<2p$, étant devenus nuls. Au prix de calculs assez laborieux on établit que $K_p\left(x-\frac{a_{q-1}+a_q}{2},\ H\right)$ ne change pas de signe pour $a_{q-1}\leqslant x\leqslant a_q$. En se servant du théorème de la moyenne, on a

$$r_q^p(f)=f^{(2p)}(\zeta_q^p)\int_{a_{q-1}}^{a_q} K_p\left(x-\frac{a_{q-1}+a_q}{2},\ H\right)dx, \tag{3}$$

où $a_{q-1}\leqslant\zeta_q^p\leqslant a_q$. Or

$$\int_{a_{q-1}}^{a_q} K_p\left(x-\frac{a_{q-1}+a_q}{2},\ H\right)dx=-\gamma_{2p}H^{2p+1}.$$

La sommation de (3) donne

$$I(f)-S_M^{2p}(f)=-\gamma_{2p}H^{2p}E_{2p}$$

avec

$$E_{2p}=H\sum_{q=1}^{M} f^{(2p)}(\zeta_q^p)$$

compris entre

$$(B-A)\min_{[A,\,B]} f^{(2p)}(x) \quad\text{et}\quad (B-A)\max_{[A,\,B]} f^{(2p)}x\,;$$

cette quantité est donc égale à $(B-A)f^{(2p)}(\zeta_p)$, $A\leqslant\zeta_p\leqslant B$. L'expression définitive de l'erreur est

$$I(f)-S_M^{2p}(f)=-\gamma_{2p}f^{(2p)}(\zeta_p)\,H^{2p}(B-A).$$

Citons plusieurs valeurs

$$\gamma_2=\frac{1}{12},\quad \gamma_4=-\frac{1}{720},\quad \gamma_6=\frac{1}{30\,240},$$

$$\gamma_8=-\frac{1}{1\,209\,600},\quad \gamma_{10}=\frac{1}{47\,900\,160}.$$

Dans la pratique, les formules d'Euler sont d'un emploi peu commode : on calcule, en plus de la fonction, ses dérivées.

Remplaçons dans la formule donnant $S_M^{2p}(f)$ les dérivées $f^{(2k-1)}(A)$ et $f^{(2k-1)}(B)$ par les dérivées des polynômes d'interpolation de degré l basés respectivement sur les points $a_0,\ \ldots,\ a_l$ et $a_N,\ \ldots$ $\ldots,\ a_{N-l}$. Si $l=2p-3$ et $l=2p-2$, les transformations intermédiaires aboutissent aux *formules* d'intégration numérique dues à

Gregory

$$I(f) \approx G_M^l(f),$$

$$G_M^l(f) = S_M^2(f) + H \sum_{k=1}^{l} \beta_k (\nabla^k f(a_M) - (-1)^k \Delta^k f(a_0)),$$

où

$$\beta_k = (-1)^{k+1} \int_0^1 \frac{x(x-1)\dots(x-k)}{(k+1)!}\, dx.$$

En particulier,

$$\beta_1 = -\frac{1}{12}, \quad \beta_2 = -\frac{1}{24}, \quad \beta_3 = -\frac{19}{720},$$

$$\beta_4 = -\frac{3}{160}, \quad \beta_5 = -\frac{863}{60\,480}, \quad \beta_6 = -\frac{275}{24\,192}.$$

La partie principale de l'erreur sur la formule $I(f) \approx G_M^l(f)$ est

$$\beta_{l+1} H^{l+2} (f^{(l+1)}(B) - (-1)^{l+1} f^{(l+1)}(A)).$$

PROBLÈME. Démontrer cette dernière relation en supposant valables tous les raisonnements ci-dessous.

Puisque le comportement des dérivées des fonctions à intégrer diffère en réalité selon le segment élémentaire, les programmes généraux d'intégration numérique font appel à des procédés divers avec recherche automatique du pas (v. §§ 17, 18). Aussi l'avènement des ordinateurs a-t-il restreint le domaine d'application des formules d'Euler-Gregory. Ces dernières s'appliquent cependant dans l'intégration de fonctions données par une table, lors du calcul d'intégrales indéfinies, pour résoudre les équations intégrales de Volterra, ainsi que dans certains autres problèmes où la fonction à intégrer doit se calculer essentiellement sur un réseau régulier.

§ 13. Règle d'évaluation pratique de l'erreur de Runge

La relation (11.6) et ses conséquences nous renseignent sur la grandeur de la partie principale de l'erreur. Cependant, le calcul direct de l'intégrale

$$K_m = \frac{\gamma_m}{2^{m+1}} \int_0^1 f^{(m)}(\varphi(t))\,(\varphi'(t))^{m+1}\, dt, \qquad (1)$$

caractérisant cette partie principale, s'avère assez pénible.

Dans la pratique on suit donc une autre voie.

Donnons-nous certains M_1 et M_2, $M_1 < M_2$, et intégrons deux fois sur les segments partiels $\left[\varphi\left(\frac{q-1}{M_1}\right), \varphi\left(\frac{q}{M_1}\right)\right]$ et $\left[\varphi\left(\frac{q-1}{M_2}\right), \varphi\left(\frac{q}{M_2}\right)\right]$ respectivement. Nous obtenons les valeurs approchées $S_{M_1}(f)$ et $S_{M_2}(f)$. Selon (11.6),

$$\begin{aligned} I(f) - S_{M_1}(f) &= K_m M_1^{-m} + \rho_{M_1}, \\ I(f) - S_{M_2}(f) &= K_m M_2^{-m} + \rho_{M_2} \end{aligned} \tag{2}$$

où $\rho_M = o(M^{-m})$. Soit $K_m \neq 0$. Si les relations (2) ne contenaient pas ρ_{M_1} et ρ_{M_2} on obtiendrait deux équations à deux inconnues $I(f)$ et K_m se prêtant à une détermination; dans le cas contraire, on n'obtient que leurs approximations. En éliminant $I(f)$, on définit K_m par

$$S_{M_2}(f) - S_{M_1}(f) = K_m (M_1^{-m} - M_2^{-m}) + \rho_{M_1} - \rho_{M_2}.$$

D'où

$$K_m = \frac{S_{M_2}(f) - S_{M_1}(f) + \rho_{M_2} - \rho_{M_1}}{M_1^{-m} - M_2^{-m}}$$

et l'erreur commise sur $S_{M_2}(f)$ s'écrit

$$\begin{aligned} I(f) - S_{M_2}(f) &= K_m M_2^{-m} + \rho_{M_2} = \\ &= \frac{S_{M_2}(f) - S_{M_1}(f)}{(M_2/M_1)^m - 1} + \frac{\rho_{M_2} - \rho_{M_1}}{(M_2/M_1)^m - 1} + \rho_{M_2}. \end{aligned}$$

Le premier terme

$$z_1 = \frac{S_{M_2}(f) - S_{M_1}(f)}{(M_2/M_1)^m - 1}$$

est calculable. Si c'est la partie principale de l'erreur, alors $I(f) - S_{M_2}(f) \approx z_1$, et z_1 sera considérée comme estimation de l'erreur. L'évaluation de l'erreur par la quantité z_1 porte le nom de *règle de Runge*. Afin que z_1 soit la partie principale de l'erreur $I(f) - S_{M_2}(f)$ de l'ordre de M_2^{-m}, il faut que

$$z_2 = \frac{\rho_{M_2} - \rho_{M_1}}{(M_2/M_1)^m - 1} + \rho_{M_2} = o(M_2^{-m}). \tag{3}$$

Si par exemple $M_2 \sim M_1 \lambda$, $\lambda > 1$, alors

$$z_2 = \frac{o(M_2^{-m})}{\lambda^m - 1} + o(M_2^{-m}) = o(M_2^{-m}),$$

et la condition nécessaire est satisfaite.

Si l'on possède une évaluation plus forte de ρ_M, $I(f) - S_{M_2}(f) \approx z_1$ sera satisfaite dans des hypothèses plus faibles sur la relation entre M_1 et M_2. Dans le calcul d'intégrales simples un cas typique est celui où

$$\rho_M = O(M^{-m-1}). \tag{4}$$

Alors

$$z_2 = \frac{O(M_2^{-m-1}) + O(M_2^{-m-1})}{(M_2/M_1)^m - 1} + O(M_2^{-m-1}).$$

Soit d'abord $M_2/M_1 \to 1$; d'après $(1+\varepsilon)^m - 1 \sim \varepsilon m$ avec ε tendant vers zéro, il vient

$$\left(\frac{M_2}{M_1}\right)^m - 1 \sim m\left(\frac{M_2}{M_1} - 1\right).$$

Alors

$$z_2 = O\left(\frac{1}{M_2^{m+1}\left(\frac{M_2}{M_1} - 1\right)}\right) + O\left(\frac{1}{M_2^{m+1}}\right) = O\left(\frac{1}{(M_2 - M_1)M_2^m}\right). \tag{5}$$

Pour que cette première condition vérifie (3) il suffit donc que $M_2 - M_1 \to \infty$ avec $M_2 \to \infty$.

Soit maintenant $M_2/M_1 \to \infty$. Alors

$$z_2 = O\left(\frac{1}{M_1^{m+1}\left(\left(\frac{M_2}{M_1}\right)^m - 1\right)}\right) + O\left(\frac{1}{M_2^{m+1}}\right) = O\left(\frac{1}{M_1 M_2^m}\right), \tag{6}$$

pour que (3) soit satisfaite, i.e. il suffit que $M_1 \to \infty$. En réunissant les estimations (5) et (6), on a

$$z_2 = O\left(\max\left(\frac{1}{M_1}, \frac{1}{M_2 - M_1}\right)\frac{1}{M_2^m}\right).$$

Par conséquent, pour que la règle de Runge s'applique au cas (4), il suffit que

$$M_1,\ M_2 - M_1 \gg 1. \tag{7}$$

PROBLÈME. Démontrer qu'il suffit que $M_1 \gg 1$ pour que la règle de Runge s'applique au cas $\rho_M = O(M^{-m-2})$.

Il peut arriver que K_m définie par (1) soit nulle ou très petite. Avec les M utilisés dans la pratique, le terme $K_m M^{-m}$ ne sera pas sensiblement supérieur en valeur absolue à ρ_M et nos constructions seront plutôt douteuses. Dans ce cas, en évaluant l'erreur par z_1, on aura souvent une idée fausse de celle-là.

Si z_1 est la partie principale de $K_m M_2^{-m}$ et si $z_2 = o(M_2^{-m})$, on est en droit d'écrire

$$I(f) - (S_{M_2}(f) + z_1) = o(M_2^{-m}). \tag{8}$$

i.e. $S_{M_2}(f) + z_1$ est une meilleure approximation de la valeur initiale de l'intégrale.

La règle de Runge est à la base de plusieurs procédures de calcul d'intégrales. On se donne par exemple certains M_0 et $\lambda > 1$ et on

calcule successivement $S_{M_k}(f)$ pour $M_k = [M_0\lambda^k]$, $k = 0, 1, \ldots$. Chaque valeur de $S_{M_k}(f)$, $k > 0$, permet de déterminer

$$z_1^k = \frac{S_{M_k}(f) - S_{M_{k-1}}(f)}{(M_k/M_{k-1})^m - 1}$$

correspondante et de tester la condition $|z_1^k| < \varepsilon$, ε étant la précision voulue. Dès que cette condition a lieu pour un certain $k = l$, on adopte $S_{M_l}(f)$ ou $S_{M_l}(f) + z_1^l$ pour valeur approchée de l'intégrale de départ.

Cette méthode appelle l'objection suivante: si les valeurs de $S_{M_k}(f)$ pour k divers sont calculées indépendamment l'une de l'autre, on accomplit un travail vain puisque nombre d'entre elles sont « superflues ».

La situation n'est cependant pas si catastrophique qu'elle le semble de prime abord. Le volume de travail nécessaire au calcul de chaque $S_{M_k}(f)$ est proportionnel à $M_k \sim M_0\lambda^k$ et le volume total à

$$\sum_{k=0}^{l} M_0\lambda^k = M_0 \frac{\lambda^{l+1} - 1}{\lambda - 1} \sim M_0 \frac{\lambda^{l+1}}{\lambda - 1}. \tag{9}$$

Si l'on ne considère comme « utile » que le calcul de $S_{M_l}(f)$, ce dernier travail est égal au rapport de $M_0\lambda^l$ à $M_0 \frac{\lambda^{l+1}}{\lambda - 1}$, i.e. à $1 - \lambda^{-1}$. Si l'on estime « utile » le travail accompli pour calculer $S_{M_{l-1}}(f)$, valeur nécessaire pour évaluer l'erreur, le pourcentage de travail utile est égal au rapport $M_0\lambda^l + M_0\lambda^{l-1}/M_0 \frac{\lambda^{l+1}}{\lambda - 1}$, i.e. $1 - \lambda^{-2}$.

L'argument suivant s'avère excellent pour parer toute observation éventuelle sur le volume important de travail inutile. Prenons $\lambda = 10$; la part du travail utile sera alors de 99 %. Cette manière de justifier l'efficacité d'une méthode en arguant d'une proportion élevée du travail « utile » est généralement probante, encore qu'illicite. Car la qualité d'une méthode doit se mesurer au coût total et non pas à la part du travail utile.

Il paraît que la procédure de calcul de l'intégrale que nous avons décrite plus haut n'est pas la meilleure. Il serait pourtant utile d'examiner le problème de choix du paramètre λ.

Pour que cette analyse soit possible, simplifions le problème et admettons que l'erreur $R_M(f)$ vérifie la condition $R_M(f) = K_m M^{-m}$; alors z_1^k est égal à $K_m M_k^{-m}$. Soit k le plus petit entier tel que $|K_m| (M_0\lambda^k)^{-m} < \varepsilon$. La dernière relation équivaut à l'inégalité

$$\frac{1}{m} \log_\lambda \left(\frac{|K_m|}{M_0^m \varepsilon} \right) < k.$$

D'où

$$k=\left[\frac{1}{m}\log_\lambda\left(\frac{|K_m|}{M_0^m\varepsilon}\right)\right]+1.$$

Adoptons une écriture plus commode

$$k=1+\frac{1}{m}\log_\lambda\left(\frac{|K_m|}{M_0^m\varepsilon}\right)-\left\{\log_\lambda\sqrt[m]{\frac{|K_m|}{M_0^m\varepsilon}}\right\}$$

avec, comme toujours, $\{y\}=y-[y]$. Le volume total de calculs est proportionnel, selon (9), à

$$\frac{M_0\lambda^k}{1-\frac{1}{\lambda}}=\frac{M_0\lambda^{\frac{1}{m}\log_\lambda\left(\frac{|K_m|}{M_0^m\varepsilon}\right)+1-\left\{\log_\lambda\sqrt[m]{\frac{|K_m|}{M_0^m\varepsilon}}\right\}}}{1-\frac{1}{\lambda}}=$$

$$=\frac{\left(\frac{|K_m|}{\varepsilon}\right)^{\frac{1}{m}}\lambda^{1-\left\{\log_\lambda\sqrt[m]{\frac{|K_m|}{M_0^m\varepsilon}}\right\}}}{1-\frac{1}{\lambda}}.$$

Le second facteur

$$g(\lambda)=\frac{\lambda^{1-\left\{\log_\lambda\sqrt[m]{\frac{|K_m|}{M_0^m\varepsilon}}\right\}}}{1-\frac{1}{\lambda}}$$

dépend de λ; puisqu'on ne sait rien sur $|K_m|$, il convient de préciser la position du problème.

On peut le faire:

1) en rendant

$$\max_{|K_m|} g(\lambda)=g_1(\lambda)=\frac{\lambda}{1-\frac{1}{\lambda}}$$

minimal en λ. Une étude directe montre que $g_1(\lambda)$ prend la valeur minimale 4 pour $\lambda=2$;

2) en étudiant la statistique du comportement de $g(\lambda)$.

La statistique établit que *si l'on se donne un* $\lambda>1$ *quelconque et qu'on obtienne de diverses sources indépendantes certains* y, *les quantités* $\{\log_\lambda y\}$ *auront une répartition asymptotiquement uniforme sur* [0, 1].

La quantité

$$y=\left\{\log_\lambda\sqrt[m]{\frac{|K_m|}{M_0^m\varepsilon}}\right\}$$

s'écrit

$$\{\log_\lambda\sqrt[m]{|K_m|}-\log_\lambda M_0-\log_\lambda\sqrt[m]{\varepsilon}\}.$$

La précision ε ne saurait être considérée comme aléatoire, puisque d'habitude elle est donnée sous une forme bien déterminée, par exemple $\varepsilon=2^{-s}$, $\varepsilon=10^{-s}$, ou $\varepsilon=0{,}5\cdot10^{-s}$. Il en est également de la valeur M_0. Par contre, $\sqrt[m]{|K_m|}$

peut très bien être aléatoire, puisque K_m caractérise la partie principale de l'erreur commise dans divers problèmes. On supposera donc naturellement que la répartition de y sur [0, 1] suit la loi uniforme. Alors

$$g_2(\lambda) = \mathrm{E}(g(\lambda)) = \int_0^1 \frac{\lambda^{1-y}}{1-\frac{1}{\lambda}}\,dy = \frac{\lambda}{\ln \lambda}.$$

On vérifie directement que la plus petite valeur de $g_2(\lambda)$, à savoir e, est atteinte pour $\lambda = e$.

Quel procédé est à préférer? Si l'on a beaucoup de problèmes à résoudre, les autres conditions étant les mêmes, l'approche statistique semble plus rationnelle. On aura alors le tableau

λ	1,5	2	2,718	3	4	8	16
$g_1(\lambda)$	4,5	4	4,300	4,5	5,333	9,143	17,667
$g_2(\lambda)$	3,699	2,886	2,718	2,730	2,886	3,848	5,772

D'où

$$\frac{g_1(e)}{g_1(2)} = 1{,}075, \qquad \frac{g_2(2)}{g_2(e)} = 1{,}062.$$

Ainsi, la substitution au point d'extrémum d'une fonction $g_i(\lambda)$ du point d'extrémum d'une autre change peu les valeurs des fonctions. On préférera $\lambda = 2$ ou $\lambda = e$, encore que, dans notre cas, le choix définitif soit sans importance: bien que $\lambda = e$ soit statistiquement meilleur, $\lambda = 2$ est traditionnel et présente certains avantages en ce qui concerne l'accumulation de l'erreur de calcul. Il est d'usage, lorsqu'on calcule des intégrales s-uples et que l'on passe à un grand nombre de points, de multiplier ce dernier par 2^s, ce qui correspond à $\lambda = 2^s$. Si $s = 1, 2$, ce choix est satisfaisant, et si $s > 2$, il est moins heureux.

La règle de Runge admet une autre utilisation. Intégrons sur un nombre de points M_1 et M_2 peu important, puis déterminons K_m. Dans le cas unidimensionnel, on a ordinairement $M_2/M_1 = 2$. A l'aide de l'équation $|K_m| M^{-m} = \varepsilon$, on définit le nombre inconnu $M = M_3$ et on intègre sur M_3 segments partiels. Vérifions le résultat par la règle de Runge (si elle est licite) en nous appuyant sur les calculs à M_2 et M_3 points. Si, pour M_2 et M_3 donnés, la règle en question s'avère douteuse, on prend un autre nombre M_4 de segments élémentaires pour déterminer l'erreur. Si la précision ne nous satisfait pas, nous nous servons de S_{M_3} et S_{M_4} pour calculer un nouveau K_m, puis le nombre cherché de segments $M = M_5$. En choisissant la stratégie optimale de cet algorithme, il faut voir s'il n'est pas possible d'éviter de nombreux calculs répétitifs en prenant M plus grand que ne le donne la solution de $|K_m| M^{-m} = \varepsilon$.

Formulons une règle gouvernant le choix du critère d'évaluation pratique de l'erreur. Soient u une solution du problème, u_n les approximations obtenues, $\lim_{n\to\infty} u_n = u$. On prend pour erreur sur

la solution approchée un certain v_n tel que

$$\frac{|v_n-(u_n-u)|}{|u_n-u|}\to 0$$

pour n tendant vers l'infini. Evidemment $|v_n| \sim |u_n - u|$.

Dans notre problème d'intégration numérique, nous adoptons comme mesure de l'erreur vérifiant cette condition la quantité z_1 si $z_2 = o(z_1)$. Or, la relation $I(f) - S_M(f) \sim z_1$ est asymptotique. Aussi, dans les cas particulièrement importants, vérifie-t-on si cette relation a lieu. Par exemple, pour le cas décrit du calcul successif des $S_{M_i}(f)$ pour $M_i = M_0\lambda^i$, on a

$$S_{M_i}(f)-S_{M_{i-1}}(f)=(I(f)-S_{M_{i-1}}(f))-(I(f)-S_{M_i}(f))=$$
$$=K_m(\lambda^m-1)M_i^{-m}+o(M_i^{-m}).$$

Si le nombre de points est tellement élevé que $o(M_i^{-m})$ est sensiblement inférieur à $K_m(\lambda^m - 1)M_i^{-m}$, alors

$$\omega_i=\frac{S_{M_{i+1}}(f)-S_{M_i}(f)}{S_{M_i}(f)-S_{M_{i-1}}(f)}\approx\frac{K_m(\lambda^m-1)M_{i+1}^{-m}}{K_m(\lambda^m-1)M_i^{-m}}=\lambda^{-m}.$$

Si ω_i est suffisamment proche de λ^{-m}, on se borne à z_1 pour l'évaluation de l'erreur.

Une fois on a eu à résoudre un problème avec une grande précision, laquelle devait être garantie. Ce problème étant insoluble avec les programmes usuels et on a fait appel à la formule de quadrature d'interpolation (voir § 1) correspondant à $m = 6$ avec division du segment d'intégration. L'intégrale se calculait successivement pour les segments partiels au nombre de $M_k = 2^k$ et on a évalué l'erreur par z_1^k, pourtant la précision voulue n'était pas atteinte même pour un très grand nombre de points. On a testé l'évaluation asymptotique par le comportement des ω_i. On a constaté que ces dernières quantités tendaient vite vers la valeur 1/4. On s'est alors demandé si le programme ne renfermait pas par erreur, au lieu d'une formule correspondant à l'ordre de convergence $m = 6$, une formule de quadrature associée à $m = 2$; c'est ce qui a eu effectivement lieu.

Dans un autre cas semblable, les quantités ω_i s'accumulaient autour de 1/3 au lieu de tendre vers 2^{-6}. Toute tentative de détecter une erreur de programme fut vaine. Il s'est trouvé que l'intégrande était de la forme $g(x)\sqrt{x+\delta}$ avec $\delta > 0$ très petit et $g(x)$ régulière. Pour un partage équidistant du segment d'intégration, l'erreur se comporte en effet comme M^{-6} à condition que M soit très grand. Pour des M réels elle se comporte comme $M^{-3/2}$ et, par conséquent, les quantités ω doivent être proches de $2^{-3/2}$. La vérification de la règle de Runge selon le critère $\omega_i \to 2^{-m}$ a été en l'occurrence fort utile: l'estimation d'erreur d'après cette règle a été d'environ 20 fois inférieure à l'erreur réelle, et n'était la vérification, le résultat définitif aurait été entaché d'une erreur inadmissible.

Voici un exemple d'emploi heureux de la règle de Runge. On a eu à trouver une importante série d'intégrales d'un même type et à garantir dans chaque cas la petitesse de l'erreur. On a décidé de faire appel à la première des procédures décrites ci-dessus avec évaluation de l'erreur par la règle de Runge. Afin

de parer l'éventualité d'un malentendu analogue à ceux des cas précédents, on a choisi plusieurs dizaines d'intégrales types, effectué les calculs et donné l'ordre d'imprimer les tableaux des valeurs des ω_l. Celles-ci ayant été convergées assez bien, on a conclu à l'applicabilité de la règle de Runge dans la résolution des problèmes de la série donnée.

§ 14. Formules de Romberg

Si la fonction à intégrer est suffisamment régulière, l'erreur dans la formule de quadrature s'écrit

$$R_M(f) = \sum_{k=1}^{l} D_k(f) M^{-i_k} + r(M) \tag{1}$$

avec

$$i_1 < \ldots < i_l, \quad r(M) = o(M^{-i_l}).$$

En supposant par exemple que la dérivée $f^{(2m+2)}(x)$ est bornée, on a, selon (12.2), pour l'erreur sur la formule des trapèzes:

$$R_M(f) = -\sum_{k=1}^{m} \gamma_{2k} \left(\frac{B-A}{M}\right)^{2k} \left(f^{(2k-1)}(B) - f^{(2k-1)}(A)\right) + \\ + \left(O(B-A)\left(\frac{B-A}{M}\right)^{2m+2}\right).$$

En général, si la fonction à intégrer est régulière, alors

$$i_2 - i_1 = \ldots = i_l - i_{l-1} = s,$$

s étant égal à 1 ou 2. Admettons qu'on a trouvé $S_M(f)$ pour $M = M_0, \ldots, M_l$. Il vient les égalités

$$I(f) = S_{M_j}(f) + \sum_{k=1}^{l} D_k(f) M_j^{-i_k} + r(M_j),$$
$$j = 0, \ldots, l.$$

Réalisons une combinaison linéaire de ces relations munies de certains coefficients c_j tels que

$$\sum_{j=0}^{l} c_j = 1. \tag{2}$$

On a

$$I(f) = \sum_{j=0}^{l} c_j S_{M_j}(f) + \sum_{k=1}^{l} D_k(f) \left(\sum_{j=0}^{l} c_j M_j^{-i_k}\right) + \sum_{j=0}^{l} c_j r(M_j).$$

Supposons qu'on a

$$\sum_{j=0}^{l} c_j M_j^{-i_k} = 0 \quad \text{pour } k = 1, \ldots, l; \tag{3}$$

alors

$$I(f)=\sum_{j=0}^{l} c_j S_{M_j}(f)+\sum_{j=0}^{l} c_j r(M_j). \tag{4}$$

Si $\sum_{j=0}^{l} c_j r(M_j)$ est négligeable, on pose

$$I(f)\approx\sum_{j=0}^{l} c_j S_{M_j}(f). \tag{5}$$

C'est la *formule de Romberg*. Les relations (2), (3) forment un système de $l+1$ équations linéaires à $l+1$ inconnues, et il est donc susceptible d'une solution.

Signalons le lien existant entre la résolution de ce système et le problème d'interpolation. Supposons, par exemple, qu'on a comme pour la formule des trapèzes:

$$I(f)=S_M(f)+\sum_{k=1}^{l} D_k M^{-2k}+r(M).$$

Récrivons cette relation:

$$Q_l(M^{-2})=S_M(f)+r(M),\ \text{avec}\ Q_l(y)=I(f)-\sum_{k=1}^{l} D_k y^k. \tag{6}$$

Son examen révèle que le problème de calcul de $I(f)$ s'énonce comme suit. Soient données les valeurs approchées du polynôme $Q_l(y)$ pour $y=M_0^{-2},\ \ldots,\ M_l^{-2}$; on demande de trouver $Q_l(0)=I(f)$. Selon la formule d'interpolation de Lagrange (§ 2, ch. II),

$$Q_l(y)=\sum_{j=0}^{l} Q_l(y_j)\prod_{i\neq j}\frac{y-y_i}{y_j-y_i}.$$

D'où

$$Q_l(0)=\sum_{j=0}^{l} c_j Q_l(y_j)\ \text{avec}\ c_j=\prod_{i\neq j}\frac{y_i}{y_i-y_j},\ y_j=M_j^{-2}, \tag{7}$$

et donc

$$I(f)=Q_l(0)=\sum_{j=0}^{l} c_j Q_l(y_j)=\sum_{j=0}^{l} c_j S_{M_j}(f)+\sum_{j=0}^{l} c_j r(M_j).$$

On a abouti à la même égalité (4) et explicité en même temps les coefficients c_j.

Les approximations (5) représentent certaines formules de quadrature. Soient, pour fixer les idées, S_M la formule des trapèzes, $l = 1$, $M_1 = 2M_0$. D'après (7), on a

$$c_0 = \frac{1}{1-2^2} = -\frac{1}{3}, \quad c_1 = \frac{1}{1-\left(\frac{1}{2}\right)^2} = \frac{4}{3}.$$

Alors (5) s'écrit

$$J(f) \approx \frac{4}{3} S_{2M_0}(f) - \frac{1}{3} S_{M_0}(f) =$$

$$= H\left(\frac{1}{6} f(A) + \frac{4}{6} f\left(A + \frac{H}{2}\right) + \frac{2}{6} f(A+H) + \ldots + \frac{1}{6} f(B)\right) \qquad (8)$$

avec $H = (B-A)/M_0$; le second membre représente donc la formule de Simpson. Faisons $l = 2$, $M_1 = 2M_0$, $M_2 = 4M_0$. Alors $c_0 = \frac{1}{45}$, $c_1 = -\frac{4}{9}$, $c_2 = \frac{64}{45}$,

$$J(f) = \frac{64}{45} S_{4M_0}(f) - \frac{20}{45} S_{2M_0}(f) + \frac{1}{45} S_{M_0}(f) =$$

$$= H\left(\frac{7}{90} f(A) + \frac{32}{90} f\left(A + \frac{1}{4}H\right) + \frac{12}{90} f\left(A + \frac{2}{4}H\right) + \right.$$

$$\left. + \frac{32}{90} f\left(A + \frac{3}{4}H\right) + \frac{14}{90} f(A+H) + \frac{32}{90} f\left(A + \frac{5}{4}H\right) + \ldots + \frac{7}{90} f(B)\right). \qquad (9)$$

Cette formule dont l'ordre de précision est 6 coïncide avec la formule de quadrature (6.4) pour $a_q - a_{q-1} \equiv H$ si le calcul des intégrales sur les segments élémentaires se fait d'après (2.10). De même, on obtient (9) en précisant la formule de Simpson avec $M_1 = 2M_0$ segments de division par la règle de Runge pour $M_2 = 2M_1$.

La règle de Romberg procure certains avantages par rapport aux formules composites. Supposons qu'on s'est donné une formule de quadrature, puis qu'on a calculé et imprimé les valeurs $S_{M_0 2^i}(f)$, $i = 0, \ldots, l$, sans réaliser la précision désirée. On essaye alors d'approcher l'intégrale en appliquant la règle de Romberg pour un certain ensemble de valeurs $S_{M_0 2^s}(f), \ldots, S_{M_0 2^p}(f)$, puis de tester la précision en l'appliquant pour l'ensemble

$$S_{M_0 2^{s+1}}(f), \ldots, S_{M_0 2^{p+1}}(f).$$

Le programme usuel le plus répandu de la méthode de Romberg travaille de façon suivante : on se donne un certain M_0 et on calcule successivement les valeurs approchées de l'intégrale par la formule des trapèzes $I(f) \approx S^{(1)}_{M_k}(f)$ pour M_k segments partiels : $M_k = M_0 \cdot 2^k$. On se servira de préférence de la formule

$$S^{(1)}_{M_k}(f) = \frac{1}{2} S^{(1)}_{M_{k-1}}(f) + \frac{B-A}{M_k} \sum_{j=1}^{M_{k-1}} f\left(A + \frac{2j-1}{M_k}(B-A)\right).$$

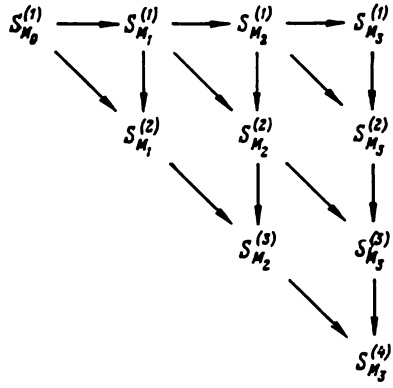
Pour chaque k, le calcul de $S^{(1)}_{M_k}(f)$ est suivi de celui de

$$S^{(2)}_{M_k}(f),\ \ldots,\ S^{(k+1)}_{M_k}(f)$$

selon la formule de récurrence

$$S^{(i)}_{M_k}(f) = S^{(i-1)}_{M_k}(f) + \frac{1}{2^i - 1}\left(S^{(i-1)}_{M_k}(f) - S^{(i-1)}_{M_{k-1}}(f)\right).$$

L'ordre des calculs est donc déterminé par le schéma



On poursuit le calcul de $S^{(i)}_{M_k}(f)$ tant que

$$\min_i |S^{(i)}_{M_k}(f) - S^{(i)}_{M_{k-1}}(f)| < \varepsilon$$

pour un certain k.

PROBLÈME. Montrer que $S^{(i)}_{M_k}(f)$ est le résultat de l'application de la règle de Romberg à

$$S^{(1)}_{M_k}(f),\ \ldots,\ S^{(1)}_{M_{k-i+1}}(f),$$

et, en particulier, que $S^{(2)}_{M_k}(f)$ coïncide avec le résultat de la formule de Simpson et $S^{(3)}_{M_k}(f)$ avec celui de (9).

§ 15. Expériences et leur discussion

Considérons un exemple d'application des résultats du § 14. Lors de la résolution d'un problème on a calculé successivement certaines approximations S_{M_k} avec $M_k = 2^k$. On a évalué l'erreur sur la solution par l'écart entre deux approximations successives $|S_{M_k} - S_{M_{k-1}}|$. Il s'est avéré que cette différence tendait très lentement vers zéro et que l'ordre de convergence était en réalité

$O(M^{-1})$. Or, les conditions impliquaient l'emploi d'une méthode avec une erreur en $O(M^{-2})$. Ces conditions sont les suivantes: on demande de résoudre une grande série de problèmes pour une contrainte donnée sur le temps de résolution; cela détermine le nombre maximal d'opérations qu'on admet pour calculer un problème quelconque de cette série. Une amélioration de la précision par la règle de Runge s'est imposée tout naturellement qui a coïncidé avec celle d'après Romberg pour $l = 1$ en supposant

$$R_M(f) \sim D_1 M^{-1} + O(M^{-2}).$$

La relation (14.5) prend alors la forme

$$I(f) \approx S^{(1)}_{M_k} = 2S_{M_k} - S_{M_{k-1}}. \tag{1}$$

On a évalué l'erreur sur la solution par la différence de deux approximations successives $|S^{(1)}_{M_k} - S^{(1)}_{M_{k-1}}|$. Cette erreur tendait elle aussi vers 0 comme M^{-1} de sorte qu'on n'a pas amélioré la convergence.

Il a fallu donc effectuer une analyse théorique plus profonde de l'erreur. Celle-là a montré que l'erreur sur une solution approchée se comportait qualitativement comme l'erreur commise lors du calcul de l'intégrale $\int_0^1 f(x)\,dx$ par la formule de quadrature

$$S_M(f) = \frac{1}{M} \sum_{q=1}^{M} f\left(\frac{q}{M}\right)$$

pour $f(x) = g(x) \ln x$; $g(x)$ est régulière, $g(0) \neq 0$. Autrement dit, on prend pour modèle le problème de calcul de cette intégrale.

On montre que dans ce cas

$$I(f) - S_M(f) = \frac{D_1 \log_2 M}{M} + \frac{D_2}{M} + O\left(\frac{\log_2 M}{M^2}\right).$$

En substituant $M_k = 2^k$, il vient

$$S_{M_k}(f) - S_{M_{k-1}}(f) = \frac{D_1 \log_2 M_{k-1}}{M_{k-1}} + \frac{D_2}{M_{k-1}} - \left(\frac{D_1 \log_2 M_k}{M_k} + \frac{D_2}{M_k}\right) +$$

$$+ O\left(\frac{k}{2^{2k}}\right) = \frac{\left(D_1 \log_2 \frac{M_k}{4} + D_2\right)}{M_k} + O\left(\frac{\log_2 M_k}{M_k^2}\right).$$

Cette relation correspond à la décroissance expérimentale de $S_{M_k} - S_{M_{k-1}}$ comme $1/M_k$; le facteur $\log_2 M_k$ varie beaucoup plus lentement que cette dernière quantité et on le met en évidence très difficilement. Formons

$$S^{(1)}_{M_k}(f) = 2S_{M_k}(f) - S_{M_{k-1}}(f);$$

on a

$$S^{(1)}_{M_k}(f) = I(f) - \frac{2D_1 \log_2 M_k}{M_k} - \frac{2D_2}{M_k} + \frac{D_1 \log_2 M_{k-1}}{M_{k-1}} + \frac{D_2}{M_{k-1}} + \\ + O\left(\frac{\log_2 M_k}{M_k^2}\right) = I(f) - \frac{2D_1}{M_k} + O\left(\frac{\log_2 M_k}{M_k^2}\right).$$

Donc,

$$I(f) - S^{(1)}_{M_k}(f) = \frac{2D_1}{M_k} + O\left(\frac{\log_2 M_k}{M_k^2}\right).$$

Les approximations $S^{(1)}_M(f)$ vérifient la relation (14.1) lorsque $l = 1$. On a toutes les raisons d'employer la règle de Runge-Romberg. Formons les approximations

$$S^{(2)}_{M_k}(f) = 2S^{(1)}_{M_k}(f) - S^{(1)}_{M_{k-1}}(f) = 4S_{M_k}(f) - 4S_{M_{k-1}}(f) + S_{M_{k-2}}(f).$$

On a

$$I(f) - S^{(2)}_{M_k}(f) = O\left(\frac{\log_2 M_k}{M_k^2}\right).$$

On obtiendrait la même $S^{(3)}_{M_k}(f)$ de la façon suivante: écrivons

$$I(f) = S_{M_i}(f) + \frac{D_1 \log_2 M_i}{M_i} + \frac{D_2}{M_i} + r(M_i), \quad i = k-2,\ k-1,\ k.$$

Choisissons convenablement les c_j et réalisons une combinaison linéaire de ces égalités

$$I(f) = \sum_{j=0}^{2} c_j S_{M_{k-j}}(f) + \sum_{j=0}^{2} c_j r(M_{k-j}).$$

L examen théorique montre qu'en calculant l'intégrale considérée on vérifie la relation

$$R_M(f) = \frac{D_1 \log_2 M}{M} + \frac{D_2}{M} + \frac{D_3 \log_2 M}{M^2} + \frac{D_4}{M^2} + O\left(\frac{\log_2 M}{M^3}\right). \quad (2)$$

Il vient à l'idée d'utiliser (2) afin d'améliorer l'ordre de précision des formules de quadrature. Néanmoins, avant de recourir à cette relation asymptotique, il serait bien de vérifier si elle a lieu avec une précision permise pour des valeurs acceptables de M.

Deux voies s'ouvrent devant nous en l'occurrence.

On pourrait essayer de choisir un problème modèle de solution connue, proche du problème considéré. Les calculs faits, on disposera de la valeur approchée de S_M et d'une erreur égale à

$$R_M = I - S_M.$$

Il se peut qu'on sache quelque chose sur le comportement de cette quantité, par exemple que

$$R_M \sim \text{const} \cdot M^{-m}. \quad (3)$$

On calcule alors, pour une suite de M_k, les valeurs $M_k^m R_{M_k}$ et on vérifie si celles-ci se stabilisent avec la croissance de M. En l'absence de toute hypothèse sur le comportement de l'erreur dans le problème donné, on procède comme suit.

Prenons le plan de coordonnées $\ln M$, $\ln\left(\frac{1}{|R|}\right)$ (fig. 3. 15.1) et portons les points $\left(\ln M_k, \ln \frac{1}{|R_{M_k}|}\right)$. Si ces points sont disposés en désordre, les nombres M_k ne sont pas grands au point de pouvoir

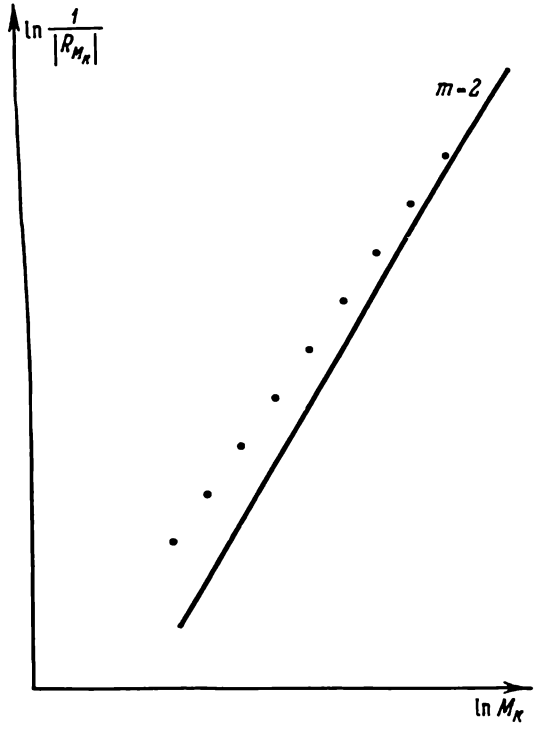


Fig. 3.15.1

isoler la partie principale de l'erreur. Supposons que dans le domaine donné de variation du paramètre M, l'égalité asymptotique (3) est réalisée avec une grande précision. Cette égalité entraîne

$$\ln\left|\frac{1}{R_M}\right| \sim m \ln M;$$

après dérivation il vient

$$\frac{d \ln \left| \frac{1}{R_M} \right|}{d \ln M} \sim m. \tag{4}$$

Notons que l'opération de dérivation des égalités asymptotiques n'est pas en général licite.

Selon (4) et pour (3) suffisamment rigoureuse, les points $\left(\ln M_k, \ln \frac{1}{|R_{M_k}|} \right)$, obtenus par calcul sur ordinateur, doivent se placer sur une courbe régulière dont la pente tend vers m. La proximité de la pente de cette courbe et d'une constante caractérise la précision de l'égalité asymptotique $R_M \sim \text{const} \cdot M^{-m}$. Si la pente varie brusquement, la règle de Runge est à exclure. En particulier, les cas

$$R_M \sim \text{const} \cdot \frac{\ln M}{M} \quad \text{et} \quad R_M \sim \text{const} \cdot \frac{1}{M}$$

se confondent pratiquement dans cette méthode car les deux fois $\frac{d \ln \left| \frac{1}{R_M} \right|}{d \ln M} \to 1$ pour $M \to \infty$.

Il existe un autre procédé permettant de vérifier nos suppositions sur l'allure de l'erreur. Si

$$R_M \sim \frac{c}{M^m} \tag{5}$$

est valable, alors

$$S_{M\lambda} - S_M \sim \frac{c\,(1-\lambda^{-m})}{M^m}. \tag{6}$$

D'autre part, si (6) est satisfaite pour $M \geqslant M_0$ et $S_M \to I$, il en est de même de (5). Aussi, au lieu de vérifier la validité pratique de (5), on teste (6), notamment en étudiant les courbes représentatives des fonctions $g(M) = (S_{M\lambda} - S_M) M^m$ ou la disposition des points

$$\left(\ln M_k, \ln \frac{1}{|S_{M_k\lambda} - S_{M_k}|} \right). \tag{7}$$

Dans le problème considéré, les points (7) se situaient, pour $\lambda = 2$, sur une courbe régulière de pente ≈ 1. On en a conclu à l'ordre de convergence (1 en l'occurrence) de la méthode en question. Pour les valeurs améliorées $S_M^{(1)}$, les points

$$\left(\ln M, \ln \frac{1}{|S_{M\lambda}^{(1)} - S_M^{(1)}|} \right) \tag{8}$$

présentaient la même disposition.

Ce comportement curieux de l'erreur a incité à s'adresser à un problème modèle et à construire une nouvelle méthode $I(f) \approx S_M^{(2)}(f)$. Les points (7) et (8) appartenant à des courbes régulières, on avait quelques raisons de supposer qu'en isolant puis en éliminant la partie principale de l'erreur par le procédé indiqué on obtiendrait une meilleure précision. Pour escompter une précision plus bonne en isolant et en éliminant les termes $\frac{D_3 \log_2 M}{M^3}$ et $\frac{D_4}{M^3}$ de la partie principale de l'erreur dans la formule $I(f) \approx S_M^{(2)}(f)$, on doit être sûr qu'il s'agit bien des termes principaux pour les M utilisés. A cet effet, on a construit l'ensemble de points

$$\left(\ln M,\ \ln \frac{1}{|S_{M\lambda}^{(2)} - S_M^{(2)}|}\right).$$

Pour des valeurs de M réalisables dans les conditions données, on n'a constaté aucune « régularité » et aucun ordre dans la disposition de ces points. D'où l'impossibilité d'améliorer la précision pour la méthode en question.

Les fig. 3.15.2-3.15.5 donnent l'erreur en fonction du nombre de points dans les différentes méthodes de calcul d'intégrales des fonctions présentant des singularités typiques. L'intégration numérique était exécutée à l'aide des programmes généraux d'intégration avec recherche automatique du pas (voir § 18). N est le nombre total de points de base de l'intégration et R_N l'erreur. Les courbes M_k correspondent aux cas où les intégrales sur les segments élémentaires ont été calculées par les formules de Markov à k points ; la courbe correspondant à la formule de Markov à trois points qui coïncide avec la formule de Simpson est notée S ; on intègre sur [0, 1].

L'examen de la fig. 3.15.4 nous révèle d'une façon particulièrement spectaculaire que les formules très rigoureuses sont parfois moins efficaces que les formules peu précises. On voit également sur les figures que pour des N petits, les courbes de l'erreur peuvent avoir une allure fort complexe et ne pas représenter des fonctions régulières ; dans ce cas l'application de la règle de Runge en vue d'évaluer pratiquement l'erreur paraît injustifiée.

Nous voudrions attirer l'attention du lecteur sur un fait d'une portée pratique. Si la solution du problème renferme des singularités non étudiées qui détériorent la convergence des méthodes, on dégagera un problème modèle simple présentant les mêmes singularités et, sur ce modèle, on choisira une méthode et testera les différents critères asymptotiques. Ce procédé permet de saisir plus rapidement le fond du problème et évite de nombreux tâtonnements sur le problème originel. En particulier, on économise ici sur la programmation et le temps machine, et l'étude est simplifiée grâce à la construction de courbes rendant mieux compte du comportement de l'erreur (on obtient plus de points (7) et (8), car c'est plus facile pour un problème simple).

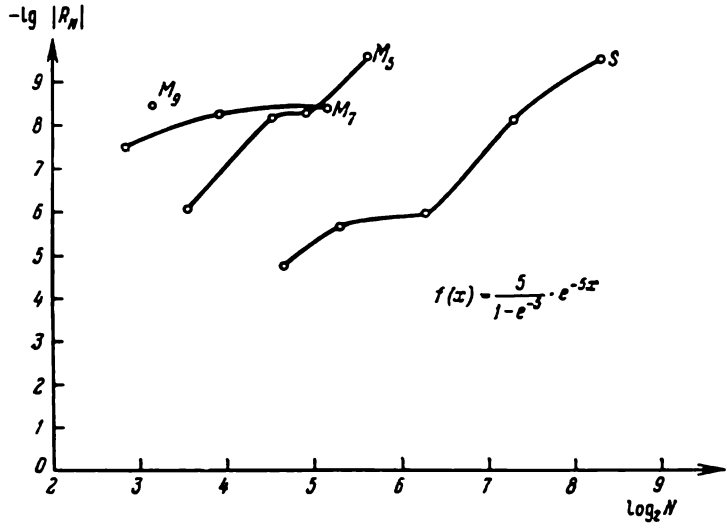


Fig. 3.15.2

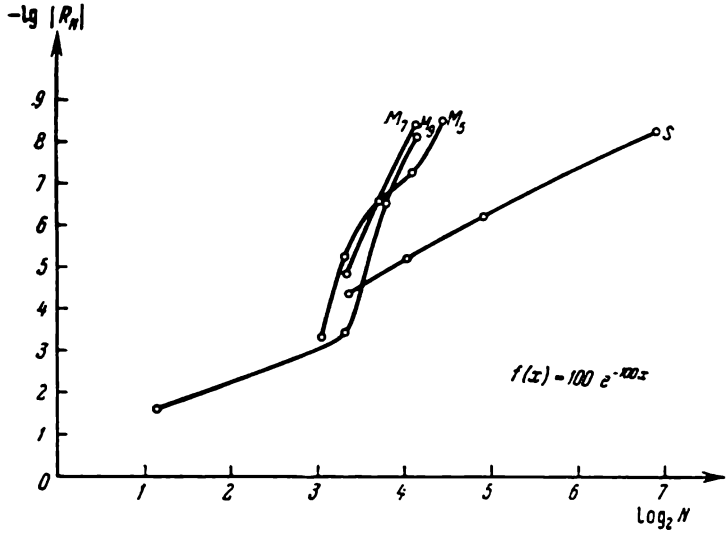


Fig. 3.15.3

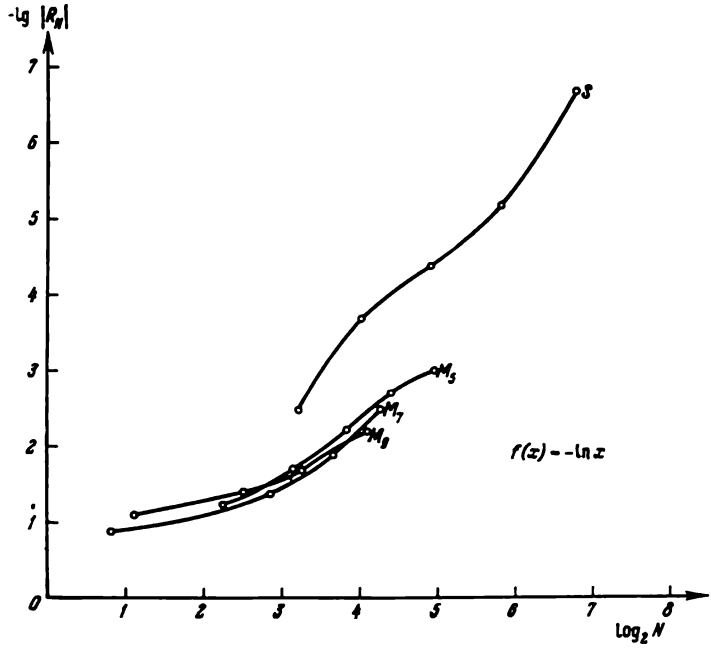


Fig. 3.15.4

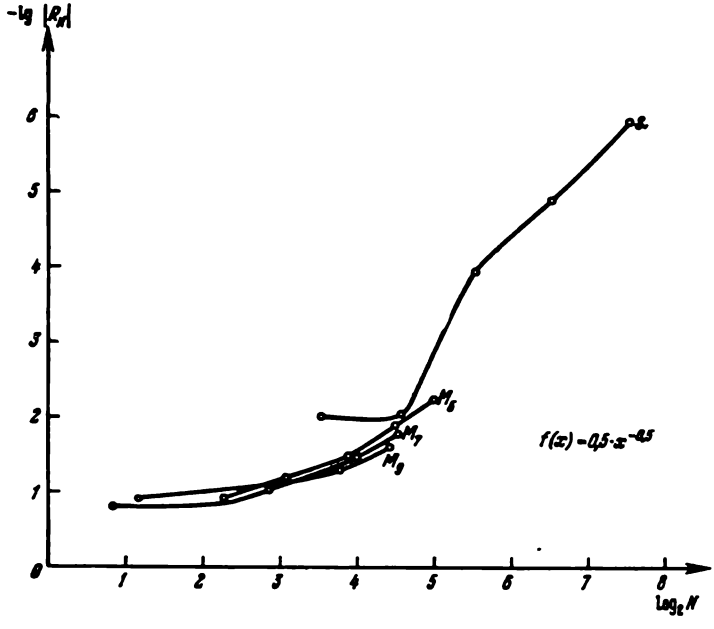


Fig. 3.15.5

§ 16. Calcul d'intégrales dans des cas non réguliers

La plupart des fonctions à intégrer qu'on rencontre en pratique sont des fonctions qui présentent des singularités (ou bien telles que leurs dérivées présentent des singularités) ou des fonctions à dérivées très grandes. Si ce comportement irrégulier de la fonction n'est pas dû à son caractère oscillatoire, les programmes généraux d'intégration avec recherche automatique du pas (voir § 17) s'avèrent assez efficaces. Dans le cas d'une faible série d'intégrales à singularités, ils permettent de résoudre le problème avec un coût minimal. Par contre, le calcul d'une grande série de telles intégrales doit faire appel à des chercheurs plus qualifiés. Décrivons plusieurs techniques susceptibles de nous aider.

1. *Dégagement de la fonction de poids.* Soit à trouver l'intégrale $\int_a^b f(x)\,dx$, avec a et b éventuellement infinis. Mettons la fonction à intégrer sous la forme $f(x) = g(x)\,p(x)$, $p(x)$ étant suffisamment élémentaire et $g(x)$ régulière. Appliquons un quelconque des procédés examinés plus haut pour calculer une intégrale avec poids. Examinons quelques exemples.

Soit à calculer $\int_{-1}^{1} \frac{dx}{\sqrt{1-x^4}}$. Représentons (x) par $\frac{1}{\sqrt{1+x^2}} \cdot \frac{1}{\sqrt{1-x^2}}$; la fonction $\frac{1}{\sqrt{1+x^2}}$ est régulière. $\frac{1}{\sqrt{1-x^2}}$ peut être considérée comme une fonction de poids à laquelle correspond la formule de quadrature d'Hermite.

Soit à calculer $\int_0^1 f(x)\,dx$, $f(x)$ pouvant s'écrire $g(x)\,x^\alpha$, avec $-1 < \alpha < 1$, $g(x)$ régulière, $g(0) \neq 0$. En calculant cette intégrale par la formule des trapèzes à pas constant $H = M^{-1}$, l'erreur tend vers zéro plus lentement que M^{-2}. Un procédé possible est le recours aux formules de Gauss associées à la fonction de poids donnée.

2. On peut aussi décomposer l'intégrale et en calculer chaque partie par les constructions du § 1. Mettons l'intégrale sous la forme

$$I = \sum_{q=1}^{M} I_q, \quad I_q = \int_{(q-1)H}^{qH} g(x)\,x^\alpha\,dx, \quad H = \frac{1}{M}.$$

Remplaçons $g(x)$ par le polynôme d'interpolation

$$P_{(q)}(x) = g((q-1)H) + (x - (q-1)H)\,\frac{(g(qH) - g((q-1)H))}{H}.$$

$$\text{Il vient } I_q \approx \int_{(q-1)H}^{qH} P_{(q)}(x)\, x^{\alpha}\, dx =$$

$$= H^{\alpha+1} q^{\alpha+2} \left(\left(\frac{1-\left(1-\frac{1}{q}\right)^{\alpha+1}}{\alpha+1} - \frac{1-\left(1-\frac{1}{q}\right)^{\alpha+2}}{\alpha+2} \right) g(qH) + \right.$$

$$\left. + \left(\frac{1-\left(1-\frac{1}{q}\right)^{\alpha+2}}{\alpha+2} - \frac{1-\frac{1}{q}-\left(1-\frac{1}{q}\right)^{\alpha+2}}{\alpha+1} \right) g((q-1)H) \right). \quad (1)$$

En sommant en q les seconds membres de (1) on obtient une formule de quadrature permettant de trouver l'intégrale initiale. Dans nombre de cas on posera

$$g(qH) = (qH)^{-\alpha} f(qH)$$

et on obtiendra une formule de la forme

$$I \approx \sum_{q=0}^{N} D(q, H) f(qH). \quad (2)$$

PROBLÈME. Etablir l'estimation d'erreur

$$\text{const} \cdot \max_{[0,\,1]} |g^{s}(x)|\, M^{-s}$$

pour la formule de quadrature (2).

Ci-dessous on examinera des procédés apparemment plus simples de calcul d'intégrales présentant des singularités. La technique d'approximation d'une intégrale par les valeurs de la fonction sur un réseau fixe, en particulier régulier, offre des avantages certains, en permettant entre autres de se ramener, en résolvant des équations intégrales, à la résolution d'un système d'équations algébriques.

Certains cas importants pour les applications admettent la précision désirée dès que $g(x)$ est remplacée par une constante sur les segments élémentaires. On pose alors

$$\int_{a_{q-1}}^{a_q} g(x)\, p(x)\, dx \approx g(\zeta_q) \int_{a_{q-1}}^{a_q} p(x)\, dx,$$

et la formule de calcul de l'intégrale de départ s'écrit

$$\int_{0}^{1} g(x)\, p(x)\, dx \approx \sum_{q=1}^{M} q(\zeta_q) \int_{a_{q-1}}^{a_q} p(x)\, dx. \quad (3)$$

Soit à calculer

$$I=\int_0^1 \frac{g(x)\,b}{b^2+x^2}\,dx \text{ avec } b \text{ petit.}$$

PROBLÈME. Montrer qu'en calculant l'intégrale par la formule des trapèzes à pas constant $a_q - a_{q-1} \equiv H = M^{-1}$, l'erreur s'évalue par

$$\text{const}\cdot\min\left(\frac{1}{Mb}\cdot\frac{1}{(Mb)^2}\right). \tag{4}$$

Si $a_q - a_{q-1} \equiv H$, la formule de quadrature (3) devient

$$I \approx \sum_{q=1}^{M} g(\zeta_q)\left(\operatorname{arctg}\frac{qH}{b}\right) - \operatorname{arctg}\left(\frac{(q-1)H}{b}\right)\Big),$$

où $(q-1)H \leqslant \zeta_q \leqslant qH$.

PROBLÈME. Etablir l'estimation

$$\text{const}\cdot\max_{[0,\,1]} |g'(x)|\,M^{-1}.$$

Dans les cas examinés plus haut, les coefficients des formules de quadrature se mettent sous forme d'intégrales

$$\int_{a_{q-1}}^{a_q} P_l(x)\,p(x)\,dx,$$

$P_l(x)$ étant des polynômes. Ces intégrales peuvent s'expliciter. Pour plusieurs classes de problèmes, où l'explicitation est impossible, il serait bon de trouver ces intégrales par intégration numérique. Ce travail supplémentaire porte ses fruits si les formules obtenues sont utilisées à maintes reprises, par exemple lors du calcul d'une série importante d'intégrales, des intégrales multiples en tant qu'intégrales répétées (voir ch. V), dans la résolution d'équations intégrales.

3. Soit à calculer

$$I_\omega(f) = \int_0^1 f(x)\exp(i\omega x)\,dx$$

avec ω grand et $f(x)$ suffisamment régulière; $\exp(i\omega x)$ sera pris pour fonction de poids. Récrivons l'intégrale comme suit:

$$I=\sum_{q=1}^{M} I_q,\quad I_q=\int_{(q-1)H}^{qH} f(x)\exp(i\omega x)\,dx,$$

et appliquons la formule du type (5.2) pour trouver I_q.

PROBLÈME. Dans cette méthode établir l'estimation

$$D_n\cdot\max_{[0,\,1]} |f^{(n)}(x)|\,M^{-n}$$

avec D_n indépendant de ω.

4. Dans certains cas la fonction à intégrer se prête à la représentation $f(x) = G(x) + g(x)$, $\int_A^B G(x)\,dx$ étant pris sous forme explicite et $g(x)$ régulière. On demande de calculer

$$I = \int_0^1 f(x)\,dx, \quad f(x) = \frac{\ln x}{1+x^2}. \tag{5}$$

Posons $G(x) = \ln x$. Alors

$$g(x) = -\frac{x^2 \ln x}{1+x^2};$$

$\int_0^1 |g''(x)|\,dx$ est finie, et l'erreur commise en calculant $\int_0^1 g(x)\,dx$ par la formule des trapèzes à pas constant $a_q - a_{q-1} = M^{-1}$ est de l'ordre de $O(M^{-2})$. Pour que l'erreur dans la formule de Simpson soit en $O(M^{-4})$, on prend $G(x) = (1 - x^2)\ln x$.

Si $f(x) = (x^2 + b^2)^{-1} \exp x$, b petit, on prendra

$$G(x) = (\operatorname*{Res}_{bi} f(z))(x - bi)^{-1} + (\operatorname*{Res}_{-bi} f(z))(x + bi)^{-1};$$

l'extension du domaine d'analyticité de la fonction à intégrer qu'on a ainsi réalisée, est particulièrement efficace ((3.8.), (4.10.1)) si l'on utilise les formules de Gauss.

5. Une autre méthode pour éliminer la singularité de la fonction à intégrer est le changement de variable d'intégration. Avec la substitution $x = \varphi(t)$, $\varphi(0) = 0$, $\varphi(1) = 1$, l'intégrale initiale $I = \int_0^1 f(x)\,dx$ se met sous la forme

$$\int_0^1 g(t)\,dt, \tag{6}$$

où

$$g(t) = f(\varphi(t))\,\varphi'(t).$$

L'apparition du facteur $\varphi'(t)$ fait disparaître les singularités de la fonction en des points isolés. Effectuons la substitution $x = t^k$ dans (5); on obtient

$$I = \int_0^1 \frac{k^2 t^{k-1} \ln t}{1+t^{2k}}\,dt;$$

si $k > 2$, l'intégrale $\int_0^1 |g''(t)|\,dt$ est finie, et l'erreur résultant de la

formule des trapèzes est donc de l'ordre de $O(M^{-2})$. En même temps que k augmente l'ordre des dérivées $g^{(n)}(t)$ pour lesquelles $\int_0^1 |g^{(n)}(t)|\,dt$ est bornée; par conséquent on peut avoir recours à des formules de quadrature toujours plus strictes.

Mais pour k très grand, les dérivées de $g(t)$, bien que finies, le sont aussi; c'est pourquoi on doit respecter une certaine proportion entre k et le nombre N de points. Le danger que présente l'emploi

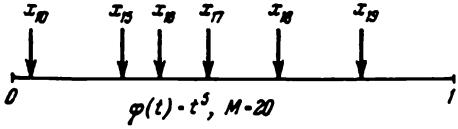


Fig. 3.16.1

des k très grands ressort ne serait-ce que de la considération suivante. Le pas constant $t_q - t_{q-1} = \frac{1}{M}$ de l'intégrale (6) correspond aux points $a_q = \varphi\left(\frac{q}{M}\right) = \left(\frac{q}{M}\right)^k$ dans l'intégrale initiale; aussi pour k élevé exploite-t-on peu de valeurs de la fonction à intégrer dans la partie droite du segment (fig. 3.16.1).

6. Nous avons vu au § 10 que l'optimisation de la répartition des points accélère elle aussi la convergence dans le calcul d'intégrales de fonctions présentant des singularités.

7. Dans certains cas, on combine quelques-unes des méthodes décrites. Soit à calculer

$$\int_0^1 g(x)\, x^{\alpha} \exp(i\omega x)\, dx,$$

ω étant grand, $g(x)$ régulière, $g(0) \neq 0$, $-1 < \alpha$. La présence du facteur $\exp(i\omega x)$ exige qu'on l'isole en tant que fonction de poids et celle de x^{α} qu'on adopte des mesures spéciales afin de pouvoir intégrer au voisinage du point $x = 0$. On ne saurait en l'occurrence effectuer le changement de variable $x = \varphi(t)$, étant donné que la fonction de poids correspondante $\exp(i\omega\varphi(t))$ exclut le calcul explicite des coefficients des formules d'interpolation. La tactique acceptable est la division du segment en parties inégales correspondant à la répartition optimale de points lors du calcul de l'intégrale de la fonction x^{α}, puis l'application à chacun des segments partiels des formules d'interpolation (5.2) associées à la fonction de poids $\exp(i\omega x)$. Si l'on a des intégrales du type $\int_0^1 g(x) x^{-\alpha} \sin \omega x\, dx$

$(2 > \alpha \geqslant 1, g(x)$ régulière, $g(0) \neq 0)$, ce procédé ne joue pas puisque la fonction non intégrable $x^{-\alpha}$ n'est pas approximée par des polynômes dans le voisinage de $x = 0$. Ici, on décompose l'intégrale originelle en $\int_0^{\varepsilon}$ et $\int_{\varepsilon}^{1}$, avec ε de l'ordre de $\frac{1}{\omega}$; la seconde intégrale se calcule par la procédure que nous venons de décrire. La fonction $\sin \omega x$ de la première intégrale n'est pas un facteur fortement oscillant, car ce choix de ε limite le nombre de ses oscillations; on trouve donc cette intégrale en adoptant, par exemple, comme répartition des points la répartition optimale pour $g(0)\, \omega x^{1-\alpha}$ qui approche la fonction à intégrer pour ωx petits.

8. Notre exposé ne serait pas complet si nous omettions la méthode de Runge. Nous l'avons déjà considérée en calculant $\int_0^1 g(x) \ln x \, dx$, i.e. l'intégrale d'une fonction présentant des singularités. L'erreur provenant de la formule des trapèzes à pas constant dans le calcul de $\int_0^1 g(x)\, x^{\alpha}\, dx$ avec $g(x)$ régulière, $g(0) \neq 0$, $-1 < \alpha < 1$, s'écrit $D_1 N^{-1-\alpha} + D_2 N^{-2-\alpha} + \ldots$; on a donc toutes les raisons pour employer la méthode de Runge.

9. On résout toute une série de problèmes en calculant les intégrales singulières du type

$$I(a) = \int_A^B \frac{g(x)}{x-a}\, dx$$

avec $a \in (A, B)$, $g(a) \neq 0$. L'intégrale est considérée au sens de la valeur principale, i.e. comme

$$\lim_{\varepsilon \to 0} \left(\int_A^{a-\varepsilon} \frac{g(x)}{x-a}\, dx + \int_{a+\varepsilon}^{B} \frac{g(x)}{x-a}\, dx \right).$$

Elle peut s'écrire comme la somme d'une intégrale sur le segment symétrique par rapport au point a et d'une intégrale de la fonction régulière sur la partie restante. L'exposé serait plus bref si nous supposons que le premier terme de la somme se transforme en $\int_{-b}^{b} \frac{g(x)}{x} dx$. Si $g(x)$ vérifie la condition de Hölder au point $x = 0$, c'est-à-dire $|g(x) - g(0)| \leqslant A\, |x|^{\alpha}$, $\alpha > 0$, le deuxième terme est égal à l'intégrale non singulière $\int_0^b \frac{g(x) - g(-x)}{x}\, dx$. En particulier, pour $g(x)$

régulière, la nouvelle fonction à intégrer $\frac{g(x)-g(-x)}{x}$ est régulière elle aussi.

Il peut arriver, par exemple lors de la résolution d'équations intégrales à noyaux singuliers, que les valeurs de la fonction $g(x)$ soient données sur un certain réseau fixe. Compte tenu de l'information disponible sur ces valeurs, on demande de calculer $I(a)$ pour divers a. Si $g(x)$ est suffisamment régulière, on opère comme suit.

Partageons $[A, B]$ en parties

$$[A_0, A_1], \ldots, [A_{M-1}, A_M], \quad A_0 = A, \quad A_M = B.$$

Sur chaque $[A_{q-1}, A_q]$ approchons $g(x)$ par le polynôme d'interpolation $L^q(x)$. Remplaçons l'intégrale initiale par la somme d'intégrales

$$i_q = \int_{A_{q-1}}^{A_q} \frac{L^q(x)}{x-a}\,dx.$$

Trouvons i_q sous forme explicite. Si $a \in (A_{q-1}, A_q)$, i_q correspondante est singulière. Si $a = A_q$, on réunit i_q et i_{q+1} (intégrales divergentes) en une intégrale singulière

$$i_q + i_{q+1} = \int_{A_{q-1}}^{A_{q+1}} \frac{\widetilde{L}^q(x)}{x-A_q}\,dx,$$

$$\widetilde{L}^q(x) = \begin{cases} L^q(x) & \text{pour } x < A_q, \\ L^{q+1}(x) & \text{pour } x > A_q. \end{cases}$$

Notons que pour que cette intégrale existe et soit singulière, on doit avoir

$$L^q(A_q) = L^q(A_{q+1}).$$

Les intégrales obtenues s'explicitent.

PROBLÈME. Soit H la longueur des segments élémentaires égaux et supposons que sur chacun d'eux $g(x)$ est approximée par interpolation linéaire. Ceci étant, l'intégrale de départ l'est par la somme

$$\sum_{q=1}^{M} \int_{A+(q-1)H}^{A+qH} \frac{g((q-1)H) + (x-(q-1)H)\,\dfrac{(g(qH)-g((q-1H))}{H}}{x-a}\,dx,$$

où $H = (B-A)\,M^{-1}$. Dans l'hypothèse de $|g''(x)|$ bornée établir l'estimation d'erreur $O(M^{-2} \ln M)$.

§ 17. Principes d'établissement des programmes usuels avec recherche automatique du pas

L'exemple 1 du § 10 montre que les intégrales de fonctions présentant des singularités du type x^b se calculent bien par les méthodes d'intégration à pas variable pour une répartition optimale des points. Il en est visiblement de même pour les fonctions

de type différent. On est donc fort tenté d'établir des programmes standard d'intégration numérique de façon que, quelle que soit la fonction, la répartition des points soit optimale ou voisine de l'optimum.

Le paragraphe 9 nous a donné une idée d'une répartition voisine de l'optimum construite en fonction du comportement de l'intégrande sur un réseau lâche. Or, s'il s'agit de fonctions à variation brusque, de celles du type x^b par exemple, la mise en pratique de cette idée aboutit à des méthodes assez médiocres. Si le réseau initial est fin, il y a une occupation mémoire importante ; s'il est lâche, on n'arrive pas à approximer de façon satisfaisante la répartition optimale des points sur les tronçons où la fonction à intégrer varie brusquement.

Aussi pour établir des programmes usuels d'intégration opte-t-on pour des procédures de répartition des points qui sont légèrement différentes et qui assurent une meilleure approximation de l'optimum dans le cas de fonctions présentant des singularités. Examinons certaines de ces procédures.

Pour calculer des intégrales sur les segments élémentaires $[a_{q-1}, a_q]$, on choisit la formule de quadrature

$$I_q(f) \approx \frac{a_q - a_{q-1}}{2} \sum_{j=1}^{m} D_j f\left(\frac{a_{q-1}+a_q}{2} + \frac{a_q - a_{q-1}}{2} d_j\right) \qquad (1)$$

et la mesure (§ 4) de l'erreur

$$\rho_q(f) = \left| \frac{a_q - a_{q-1}}{2} \sum_{j=1}^{m} B_j f\left(\frac{a_{q-1}+a_q}{2} + \frac{a_q - a_{q-1}}{2} d_j\right) \right|. \qquad (2)$$

Soit à trouver

$$\int_A^B f(x)\,dx, \quad A = a_0.$$

La première procédure qu'il semble naturel de qualifier d'*horizontale* se définit par la donnée des paramètres β, $\sigma < 1$, h_0 et ε_0. Posons $\varepsilon_1 = \varepsilon_0 \beta$. Supposons qu'on ait déjà calculé par un procédé quelconque une valeur approchée de $\int_{a_0}^{a_q} f(x)\,dx$. Le programme dispose à tout instant d'une certaine valeur h_q avec laquelle on commence à calculer la partie restante de l'intégrale. Calculons $\rho_q(f)$ correspondant au segment $[a_q, a_q + h_q]$. Si $\rho_q(f) \leqslant \varepsilon_0$, trouvons la valeur approchée de $\int_{a_q}^{a_q+h_q} f(x)\,dx$ selon la formule (1) et posons $a_{q+1} =$

$= a_q + h_q$. Nous avons obtenu la valeur approchée de $\int_{a_0}^{a_{q+1}} f(x)\,dx$. Si $\varepsilon_1 < \rho_q(f)$, faisons $h_{q+1} = h_q$ et, dans le cas contraire, $h_{q+1} = \frac{h_q}{\sigma}$. Nous sommes prêts à faire le pas suivant. Si $\rho_q(f) > \varepsilon_0$, il faut prendre σh_q pour nouvelle valeur de h_q et revenir au point de départ: on dispose de la valeur de $\int_{a_0}^{a_q} f(x)\,dx$ et on s'est donné le pas h_q. Les conditions initiales de la procédure sont

$$q = 0, \quad \int_{a_0}^{a_0} f(x)\,dx = 0, \quad h_0.$$

La procédure doit posséder un bloc de fin de travail: s'il s'avère que $a_q + h_q > B$, on pose $h_q = B - a_q$. Il est d'usage de prendre $\sigma = \frac{1}{2}$.

Une autre procédure, qu'il est logique d'appeler verticale, se définit par la donnée de ε_0 et s'exécute comme suit. Supposons qu'une nécessité se fait sentir, à un certain pas, de calculer l'intégrale sur le segment partiel $[c, d]$, soit $\int_c^d f(x)\,dx$; on trouve la quantité $\rho(f)$ associée à ce segment. Si elle est inférieure à ε_0, on utilise la formule (1) correspondante et on passe au segment suivant à droite. Dans le cas contraire, on prend pour segments élémentaires $[c, (c + d)/2]$ et $[(c + d)/2, d]$ et le programme passe au calcul de l'intégrale sur le segment à gauche. Au début le programme calcule $\int_A^B f(x)\,dx$. Un certain inconvénient de cette procédure est la nécessité de mémoriser les segments élémentaires sur lesquels on n'a pas intégré. La répartition des points réalisée par ces procédures n'est pas asymptotiquement optimale au sens du § 9 parce que $\rho_q(f)$ n'est pas le terme principal de l'estimation de l'erreur sur la formule (1), mais son estimation grossière trop forte; les segments élémentaires ne peuvent prendre que des valeurs espacées; ils sont de la forme $h_0\sigma^k$ pour la première procédure et $(B - A)\,2^{-k}$ pour la seconde.

Considérons certaines particularités de l'utilisation des procédures décrites.

Si l'on veut, dans certains cas, rapprocher la répartition des points de l'optimum, les quantités $\rho_q(f)$ correspondant aux segments partiels sont comparées non pas à ε_0, mais à $\varepsilon_0 h_q^\gamma$, où γ est choisi spécialement. Quand on a commencé à se servir des machines à cal-

culer, il était d'usage de comparer $\rho_q(f)$ à $\varepsilon_0 h_q$, i.e. de prendre γ égal à l'unité. Avec cette convention la répartition des points s'écartait notablement de l'optimum pour les fonctions variant brusquement, et pour les fonctions présentant des discontinuités, le programme n'arrivait pas à calculer l'intégrale. En effet, soit

$$f(x) = \begin{cases} 0 \text{ pour } x < c, \\ 1 \text{ pour } x > c. \end{cases}$$

Si, q étant quelconque, $\rho_q(f)$ renferme les valeurs de f aux points où $f = 0$ et $f = 1$, $\rho_q(f)$ est de l'ordre de h_q, et on n'a aucune raison de penser que

$$|\rho_q(f)| \leqslant \varepsilon_0 h_q$$

sera satisfaite pour h_q petits; le programme réalisera le fractionnement du pas jusqu'à la remise à zéro. Si la fonction varie brusquement sur un petit tronçon, le pas sera fractionné plus que de raison. Les utilisateurs en ont pris note, et au bout d'une analyse théorique ils ont convenu de poser $\gamma = 0$ si rien ne s'y opposait.

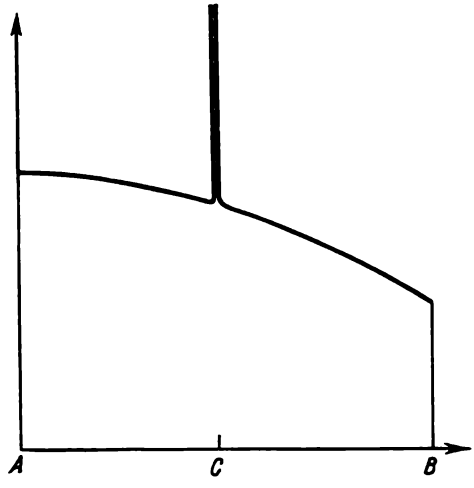


Fig. 3.17.1

En ce qui concerne les fonctions variant brusquement, il convient de noter de plus que le programme peut ne pas remarquer leur intervalle de variation. Soit $\rho(f) = 0$ si $f(x) = P_s(x)$ est un polynôme de degré s, et supposons que la fonction à intégrer réelle s'écrit

$$f(x) = P_s(x) + g(x)$$

avec $g(x)$ ne différant en fait de zéro que sur un petit intervalle, par exemple

$$g(x) = \frac{1}{2\sqrt{\pi\varepsilon}} \exp\left(\frac{(x-C)^2}{2\varepsilon}\right)$$

(fig. 3.17.1). Pour fixer les idées, adressons-nous à la première des procédures susmentionnées. Si le segment partiel $[a_{q-1}, a_q]$ est distant du point C, alors $\rho(f) \approx \rho(P_s) = 0$, et le programme augmentera le pas. A proximité du point C, le pas peut s'avérer si grand que le voisinage où $g(x)$ est sensiblement supérieure à 0 est limité par des points de base. Alors $\rho(f)$ sera proche de zéro également sur le segment contenant C. En définitive, la valeur

obtenue de $I(f)$ sera $I(P_s)$; l'erreur sur cette valeur approchée est voisine de un. Aux fins de contrôle, on aurait pu intégrer pour une autre valeur de ε_0; pourtant la probabilité est grande de répéter le résultat.

L'inconvénient dont nous venons de parler n'est pas propre aux seules méthodes d'intégration avec recherche automatique du pas. En calculant cette même intégrale par des formules à pas constant il se peut, si $H \gg \sqrt{\varepsilon}$, que $g(x) \approx 0$ en tous les points de base de l'intégration, d'où la valeur approchée $I^* \approx I(P_s)$. En effectuant l'intégration numérique pour plusieurs pas différents et en évaluant l'erreur conformément à la règle de Runge, on en conclut à tort que $I^* \approx I(P_s)$. D'une façon générale, dans les algorithmes d'intégration avec recherche automatique du pas, de pareilles conclusions erronées sont plus fréquentes.

En rédigeant des programmes standard il faut toujours tenir en équilibre les deux objectifs contraires: garantir la précision désirée pour toute fonction à intégrer ou calculer rapidement l'intégrale avec une précision voulue dans la plupart des problèmes demandant à être résolus.

S'agissant des méthodes très efficaces on peut parler d'un principe d'incertitude: la sécurité est en raison inverse de la vitesse du travail. Pour éviter que le programme « saute » les domaines de variation brusque de la fonction, on prévoit des segments avec fractionnement donné d'avance du pas. Ainsi, la description d'un programme usuel du paragraphe suivant comprend les extrémités d'un certain segment $[\alpha, \beta]$. Si le segment partiel $[a_{q-1}, a_q]$ a des points communs avec $[\alpha, \beta]$, le choix du segment partiel suivant obéit à une règle plus compliquée. On se donne en qualité de $[\alpha, \beta]$ un segment sur lequel la fonction à intégrer varie brusquement.

Faisons attention au choix de σ. Soit

$$\rho(f) = c_s \left| f^{(s)}\left(\frac{a_{q-1}+a_q}{2}\right)\right| (a_q - a_{q-1})^{s+1} + o\left((a_q - a_{q-1})^{s+1}\right) \qquad (3)$$

pour $f(x)$ régulière. Pour comprendre le mécanisme de la modification du pas, supposons que $f^{(s)}(x)$ soit constante par morceaux. On a dans son domaine de constance

$$\rho(f) = c_s \left| f^{(s)}(x)\right| (a_q - a_{q-1})^{s+1}.$$

Examinons le cas $\beta < \sigma^{s+1}$. Admettons que le pas h^0 qui est le pas initial de l'intégration dans le domaine donné vérifie $c_s \left| f^{(s)}(x)\right| (h^0)^{s+1} \leqslant \varepsilon_1$. Le pas augmentera chaque fois de σ^{-1} fois jusqu'à ce que $\rho(f)$ devienne supérieure à ε_1. Puisque pour un pas σ^{-1} fois moindre, on avait $\rho(f) \leqslant \varepsilon_1$, l'intégration aura lieu avec un pas minimal $h = h^0\sigma^k$ tel que $c_s \left| f^{(s)}(x)\right| h^{s+1} > \varepsilon_1$. De même, si le pas initial vérifie $c_s \left| f^{(s)}(x)\right| (h^0)^{s+1} > \varepsilon_0$, le pas sera le plus grand des $h^0\sigma^k$ tels que $c_s \left| f^{(s)}(x)\right| h^{s+1} \leqslant \varepsilon_0$. Sous les hypothèses faites $\left(\frac{\varepsilon_1}{\varepsilon_0} = \beta < \sigma^{s+1}\right)$, il se peut que les pas pour le présent domaine soient choisis selon le comportement de la

fonction pour x plus faibles. D'autre part, nous avons vu que le pas d'intégration doit être réparti de façon que l'estimation de l'erreur soit sensiblement identique pour tous les segments partiels. On en tire une conclusion en faveur de la valeur $\beta \geqslant \sigma^{s+1}$.

Considérons ce cas. On peut prendre un pas $h = h_0\sigma^k$ tel que

$$\varepsilon_0\sigma^{s+1} < c_s \mid f^{(s)}(x) \mid h^{s+1} \leqslant \varepsilon_0. \tag{4}$$

Si

$$\varepsilon_1 < c_s \mid f^{(s)}(x) \mid h^{s+1} \leqslant \varepsilon_0, \tag{5}$$

c'est justement le pas qu'on adoptera pour intégrer sur le segment considéré. Cependant, pour

$$\varepsilon_0\sigma^{s+1} < c_s \mid f^{(s)}(x) \mid h^{s+1} \leqslant \varepsilon_1 \tag{6}$$

le programme élaborera le pas suivant égal à h/σ. En vertu de (6), on a alors $\varepsilon_0 < \rho(f)$ et h/σ ne sera pas accepté. Le programme fera appel à un pas h qui vérifie (6) ; par conséquent, l'intégration sur le segment courant est considérée comme terminée et on prend h/σ pour pas initial sur le segment suivant. Ainsi, l'intégration utilise en fait le pas h, et à chaque pas on essaye d'intégrer avec h/σ. Si les calculs avec h et h/σ sont indépendants, on fournit un effort double à chaque pas. L'inégalité (5) est équivalente à

$$\log_{\sigma^{s+1}}\varepsilon_1 \geqslant \log_{\sigma^{s+1}} \mid f^{(s)}(x) \mid + \log_{\sigma^{s+1}}(c_s h_0^{s+1}) + k \geqslant \log_{\sigma^{s+1}}\varepsilon_0. \tag{7}$$

La quantité $\{\log_{\sigma^{s+1}} \mid f^{(s)}(x) \mid\}$ a une loi uniforme sur [0, 1] puisque c'est vrai (cf. p. 165) pour la répartition $\{\log_a g\}$ avec a quelconque et g de sources indépendantes.

On montre que la probabilité de l'événement : l'inégalité (7) est satisfaite pour un k entier, est $\log_{\sigma^{s+1}}\beta$.

Si l'inégalité (7) n'a pas lieu, alors, pour $\sigma^{s+1} \leqslant \beta$, on a (6) ; la probabilité de cette dernière inégalité est donc $1 - \log_{\sigma^{s+1}}\beta$. Ainsi, avec notre procédure, on effectue l'intégration avec le pas h déterminé par la condition (4), et l'espérance mathématique du coût par pas est proportionnelle à

$$\log_{\sigma^{s+1}}\beta + 2(1 - \log_{\sigma^{s+1}}\beta) = 2 - \log_{\sigma^{s+1}}\beta.$$

Nous avons là une expression quantitative de l'accroissement moyen du coût pour $\beta > \sigma^{s+1}$. Elle parle en faveur de la valeur $\beta = \sigma^{s+1}$. L'égalité $\rho(f) = c_s \mid f^{(s)}(x) \mid (a_q - a_{q-1})^{s+1}$, $f^{(s)}(x) \neq \text{const}$, est asymptotique. Aussi a-t-on étudié, pour certaines classes de problèmes, l'allure de $\rho(f)$ pour $a_q - a_{q-1}$ réels pas très petits. Si l'ordre de décroissance de $\rho(f)$ s'avérait différent, β changeait en conséquence.

Tout comme dans l'intégration à pas constant on désire évaluer l'erreur commise. Considérons d'abord la seconde procédure qui est d'une description plus simple. Prenons toujours une fonction constante par morceaux $f^{(s)}(x)$. Le pas d'intégration dans le domaine de constance de $f^{(s)}(x)$ est défini pour cette procédure par la condition suivante : c'est le plus grand pas h de la forme $(B - A)\,2^{-k}$ pour lequel

$$c_s \mid f^{(s)}(x) \mid h^{s+1} \leqslant \varepsilon_0. \tag{8}$$

Ce pas maximal vérifie évidemment l'inégalité

$$\varepsilon_0 2^{-(s+1)} < c_s \mid f^{(s)}(x) \mid h^{s+1}.$$

Effectuons l'intégration avec un certain $\varepsilon_0' < \varepsilon_0$. Dans le domaine où

$$\varepsilon_0' < c_s \mid f^{(s)}(x) \mid h^{s+1} \leqslant \varepsilon_0,$$

l'ancien pas d'intégration est inopérant, c'est pourquoi on le divisera en deux au moins. Dans le domaine où

$$\varepsilon_0 2^{-(s+1)} < c_s \mid f^{(s)}(x) \mid h^{s+1} \leqslant \varepsilon_0',$$

le pas restera entier. Donc, pour $\varepsilon_0 2^{-(s+1)} < \varepsilon_0'$, le partitionnement du pas n'intéressera qu'une partie du segment d'intégration ou bien il n'aura pas lieu du tout. La confrontation des résultats des calculs avec ε_0 et ε_0' ne fournira pas d'information sur l'erreur provenant de l'intégration dans le domaine où le pas est conservé. Pour

$$\varepsilon_0' \leqslant \varepsilon_0 2^{-(s+1)} \tag{9}$$

le pas est obligatoirement fractionné ; il n'en découle pas pour autant que la comparaison des résultats des calculs avec ε_0 et avec ε_0' quelconque satisfaisant à (9) garantit l'exactitude du résultat. Soit

$$\varepsilon_0 2^{-(l+1)(s+1)} < \varepsilon_0' < \varepsilon_0 2^{-l(s+1)} ;$$

dans le domaine où l'on avait antérieurement

$$\varepsilon_0' 2^{l(s+1)} < c_s \mid f^{(s)}(x) \mid h^{s+1} \leqslant \varepsilon_0, \tag{10}$$

on constate une réduction du pas de l fois ; et dans celui pour lequel on avait

$$\varepsilon_0 2^{-(s+1)} < c_s \mid f^{(s)}(x) \mid h^{s+1} \leqslant \varepsilon_0' 2^{l(s+1)} \tag{11}$$

le pas diminue de $l+1$ fois. Supposons que l'erreur d'intégration de paramètre ε_0 soit la somme de deux termes :

$$R_{\varepsilon_0}(f) = R^1_{\varepsilon_0}(f) + R^2_{\varepsilon_0}(f) ;$$

le premier terme représente la somme des erreurs commises dans le domaine vérifié par (10), et le second celles qui affectent le domaine avec (11). Dans le cas de ε_0' l'erreur prend la forme

$$R_{\varepsilon_0'}(f) = R^1_{\varepsilon_0'}(f) + R^2_{\varepsilon_0'}(f).$$

Supposons que pour $f(x)$ régulière, l'erreur sur la formule de quadrature (1) s'écrit

$$\gamma_n f^{(n)}\left(\frac{a_{q-1}+a_q}{2}\right)(a_q - a_{q-1})^{n+1} + o\left((a_q - a_{q-1})^{n+1}\right).$$

Avec toutes les conditions d'existence de la partie principale de l'erreur (§ 11), en diminuant le pas sur un certain domaine de 2^k fois, l'erreur diminue de 2^{kn} fois. Supposons respectivement que

$$R^1_{\varepsilon_0'}(f) \sim R^1_{\varepsilon_0'}(f)\, 2^{-nl}, \quad R^2_{\varepsilon_0'}(f) \sim R^2_{\varepsilon_0}(f)\, 2^{-n(l+1)}. \tag{12}$$

Alors

$$I(f) - S_{\varepsilon_0}(f) = R^1_{\varepsilon_0}(f) + R^2_{\varepsilon_0}(f), \tag{13}$$

$$I(f) - S_{\varepsilon_0'}(f) = R^1_{\varepsilon_0}(f)\, 2^{-nl} + R^2_{\varepsilon_0}(f)\, 2^{-n(l+1)}. \tag{14}$$

En éliminant l'inconnue $I(f)$, il vient

$$S_{\varepsilon_0'}(f) - S_{\varepsilon_0}(f) = R^1_{\varepsilon_0}(f)\,(1 - 2^{-nl}) + R^2_{\varepsilon_0}(1 - 2^{-n(l+1)}). \tag{15}$$

Le second membre de (15) n'est proportionnel ni au membre homologue de (13) ni à celui de (14), ce qui exclut la possibilité d'obtenir la partie principale de $R_{\varepsilon_0}(f)$ et de $R_{\varepsilon_0'}(f)$. Si nl est très élevé, le second membre de (15) est proche de $R_{\varepsilon_0}(f)$. On se demande cependant s'il vaut la peine d'utiliser cette expression pour évaluer l'erreur sur $S_{\varepsilon_0}(f)$, étant donné qu'il y a intérêt à le faire en ce qui concerne une meilleure approximation, à savoir $S_{\varepsilon_0'}(f)$. Dans le cas $\varepsilon_0' = 2^{-nl}\varepsilon_0$, il y a dégénérescence du domaine où est réalisée (11), donc $R^2_{\varepsilon_0'}(f) = 0$. Alors

$$I(f) - S_{\varepsilon_0}(f) = R^1_{\varepsilon_0}(f), \quad I(f) - S_{\varepsilon_0'}(f) = R^1_{\varepsilon_0}(f)\, 2^{-nl}$$

et l'on peut se servir de la règle de Runge.

En se livrant à une analyse complémentaire assez ardue, on constate que dans l'hypothèse de $f^{(n)}(x)$ continue et en astreignant le plus grand pas possible à diminuer d'une façon proportionnelle à $\varepsilon_0^{1/(s+1)}$, on arrive à justifier les raisonnements ci-dessus.

Revenons à la première procédure de recherche du pas. Soit $\beta = \sigma^{s+1}$, pour fixer les idées, bien que les raisonnements restent valables quel que soit β inférieur à un. En multipliant (4) par λ^{s+1}, il vient

$$\varepsilon_0 \lambda^{s+1} \sigma^{s+1} < c_s \,|\, f^{(s)}(x) \,|\, (\lambda h)^{s+1} \leqslant \varepsilon_0 \lambda^{s+1}.$$

Il en résulte qu'en multipliant simultanément ε_0 par λ^{s+1} et h_0 par λ, le pas d'intégration varie de façon proportionnelle : il se multiplie par λ. Aussi, pour toutes les hypothèses du § 11, l'erreur totale d'intégration varie-t-elle d'environ λ^n fois. Ainsi, on a

$$I(f) - S^{\lambda h_0}_{\lambda^{s+1}\varepsilon_0}(f) \approx (I(f) - S^{h_0}_{\varepsilon_0}(f))\, \lambda^n, \tag{16}$$

$S^H_\varepsilon(f)$ étant le résultat de l'intégration à pas initial H avec ε pour valeur du paramètre de la précision par pas. Nous pouvons nous servir de (16) pour évaluer l'erreur selon Runge dans l'intégration numérique avec recherche automatique du pas.

Nous avons donc obtenu, entre paramètres du schéma des relations qui sont nécessaires si l'on veut appliquer la règle de Runge pour évaluer l'erreur.

Notons qu'une variation de σ^k fois du pas initial h_0 (k entier), les autres paramètres de la méthode étant fixes, ne modifie pas sensiblement l'erreur sur le résultat : en diminuant le pas, σ fois ou en l'augmentant, $1/\sigma$ fois, le programme débouchera rapidement sur le pas vérifiant la condition

$$\varepsilon_0 2^{-(s+1)} < c_s \,|\, f^{(s)}(x) \,|\, h^{s+1} \leqslant \varepsilon_0.$$

§ 18. Programmes usuels d'intégration numérique

La description d'un programme standard comprend d'ordinaire les parties suivantes : but, appel de procédure, commentaire, description de la méthode, corps de procédure, tests.

Prenons comme exemple les descriptions partielles des procédures Algol d'intégration numérique.

La construction des programmes de calcul d'intégrales simples et multiples obéissant aux mêmes principes on se bornera à une procédure de calcul d'intégrales multiples.

1. *Procédure* INT 120. Appel de procédure: Z: INT 120 (A, B, F, P, I, X). Paramètres formels de la procédure: **reel** A (borne inférieure d'intégration); **reel** B (borne supérieure d'intégration); **reel procedure** F (fonction procédure sans paramètres donnant la fonction à intégrer $f(x)$); **entier** P (nombre conventionnel prenant la valeur 1 ou 2 suivant la formule de quadrature utilisée; si P = 1, on emploie la formule de Gauss, si P = 2, la formule de Simpson); **entier** I (nombre donnant le jeu de points de base de l'intégration; le nombre N de points est déterminé par la procédure d'après la règle suivante: $N = 2I$ dans le cas de la formule de Gauss et $N = 2I + 1$ pour la formule de Simpson. Quand P = 2, I est quelconque et supérieur à zéro, et quand P = 1, I ne prend que les valeurs suivantes: 1, 2, 3, 4, 6, 8, 12, 16, 24, 32, 48); **reel** X (argument de la fonction $f(x)$); **reel** Z (valeur calculée de l'intégrale).

Les procédures INT 110 et INT 111 servent à calculer les intégrales du type (a) et (b):

$$\text{(a)} \int_A^B f(x)\,dx, \quad \text{(b)} \int_A^B f(x)\cos(\rho x+\varphi)\,dx.$$

2. *Procédure* INT 110. Appel de procédure:

Z: = INT 110 (X, A, B, RO, FI, E, F, L).

Paramètres formels: **reel** X (argument de la fonction $f(x)$); **reel** A (borne inférieure d'intégration; si elle est égale à $-\infty$, on doit poser $A \geqslant -10^{19}$ tel que $\int_{-\infty}^{A} f(x)\,dx$ soit petite); **reel** B (borne supérieure d'intégration; si elle est infinie, on posera $B \leqslant 10^{19}$ tel que $\int_{B}^{\infty} f(x)\,dx$ soit petite; B doit être toujours supérieur à A); **reel** RO, FI (paramètres $\rho\varphi$ du facteur oscillant $\cos(\rho x + \varphi)$; dans le calcul d'intégrales du type (a) RO et FI doivent être égaux à zéro); **reel** E (précision donnée du calcul de l'intégrale; la valeur approchée sera entachée de l'erreur $E\left(1+\left|\int_A^B f(x)\,dx\right|\right)$; **reel procedure** F (fonction procédure sans paramètres donnant la fonction à intégrer $f(x)$); **etiquette** L (étiquette où l'on se renvoie si la précision donnée est impossible à atteindre ou si l'utilisateur pose A supérieur à B); **reel** Z (valeur calculée de l'intégrale).

Commentaire: 1) Si la fonction à intégrer présente une singularité en pic sur le segment d'intégration (voir fig. 3.17.1), on s'adressera à la procédure INT 111 ci-dessus.

2) E caractérise l'erreur absolue commise sur la valeur approchée de l'intégrale si cette dernière est inférieure à l'unité et l'erreur relative dans le cas contraire. Il arrive que l'utilisateur se fixe une précision trop grande impossible à atteindre pour la fonction donnée à intégrer. L'ordinateur fait alors imprimer la phrase : « la précision désirée ne peut pas être atteinte », et la procédure s'arrête à l'étiquette L.

3) L'utilisateur doit posséder son bloc étiqueté L où l'on se renvoie en cas de refus. Ainsi le bloc élémentaire Y: = 1/0 provoque l'arrêt de la machine.

3. *Procédure* INT 111. Appel de procédure : Z : = INT 111 (X, SI, A, B, RO, FI, ALFA, BETA, TAU, EPS, E, E1, F, L). *Paramètres formels non définis dans la déclaration de la procédure* INT 110 : **entier** SI (nombre conventionnel permettant le calcul des valeurs intermédiaires de l'intégrale ; les seules valeurs de SI peuvent être 0 ou 1 ; si SI = 0, la machine ne sort pas de valeurs intermédiaires) ; **reel** ALFA (origine du segment partiel sur lequel on intègre à pas donné d'avance (début de la singularité)) ; **reel** BETA (extrémité de ce segment (fin de la singularité)) ; **reel** TAU (pas donné d'avance de l'intégration sur le segment $[\alpha, \beta]$ où la fonction présente la singularité) ; **reel** EPS (erreur absolue permise de calcul de l'intégrale par pas) ; **reel** E1 (estimation (peut-être un peu forte) de l'erreur absolue).

COMMENTAIRE : 1) Si la fonction à intégrer $f(x)$ présente en un certain point du segment d'intégration une singularité en pic (voir fig. 3.17.1), on parcourt un voisinage de ce point (que l'on note $[\alpha, \beta]$) à pas donné d'avance τ (TAU). Il faut que α (ALFA) soit inférieur à β (BETA) ; s'il n'existe pas de segment partiel à parcourir avec un pas donné d'avance, les paramètres formels α, β, τ peuvent être quelconques, mais $\beta < \alpha$ nécessairement.

2) En calculant l'intégrale l'utilisateur doit se donner l'erreur absolue par pas ε (EPS) et l'erreur sur tout l'intervalle E ; si l'on ne dispose pas d'information qui pourrait dicter le choix de ε, on pose $\varepsilon = E$.

3) Si l'utilisateur se donne SI = 1, on imprimera, au cours du calcul, le nombre SI même et la valeur approchée de l'intégrale pour EPS égal à ε, $\frac{\varepsilon}{128}$, $\frac{\varepsilon}{128^2}$, . . . ; il y aura autant de résultats imprimés que de calculs sur le segment donné, et cela tant qu'on n'aboutira pas à la précision voulue E.

Description de méthodes d'intégration pour les intégrales du type (a). La droite infinie sur laquelle on intègre est partagée en deux par le point M ; on a

$$\int_{-\infty}^{\infty} f(x)\,dx = \int_{M}^{\infty} f(x)\,dx - \int_{M}^{-\infty} f(x)\,dx.$$

S'il y a singularité, $M = \alpha$, sinon $M = 0$. Voici la marche des calculs sur la demi-droite infinie. On divise celle-ci en segments Δ_i de longueur $|\Delta_1| = 2, \ldots$ $\ldots, |\Delta_n| = 2^n$. L'origine de Δ_1 coïncide avec le point M, son extrémité soit avec le point $M + 2$ (pour l'intégrale $\int_M^{\infty} f(x)\,dx$), soit avec $M - 2$ (pour l'intégrale $\int_M^{-\infty} f(x)\,dx$). Donc,

$$\int_M^{\pm\infty} f(x)\,dx = \sum_{i=1}^{\infty} \int_{\Delta_i} f(x)\,dx.$$

Sur chaque Δ_i le calcul se fait avec recherche automatique du pas conformément aux formules de Markov. Soient I_i la valeur de l'intégrale sur Δ_i et E_i l'erreur absolue de calcul. Trouvons les intégrales sur les segments successifs Δ_i, Δ_{i+1}, Δ_{i+2}. Le calcul de l'intégrale sur la demi-droite infinie s'arrête si

$$0 \leqslant \frac{|I_{i+1}|^2}{|I_i| - |I_{i+1}|} < E \quad \text{et} \quad 0 \leqslant \frac{|I_{i+2}|^2}{|I_{i+1}| - |I_{i+2}|} < E. \tag{1}$$

Soit

$$\sum_{k=1}^{i+2} E_k < E\left(1 + \left|\sum_{k=1}^{i+2} I_k\right|\right); \tag{2}$$

on en conclut que l'intégrale sur la demi-droite infinie est calculée avec la précision E donnée et qu'elle vaut $I^1 = \sum_{k=1}^{i+2} I_k$ (si l'on demande de trouver l'intégrale sur toute la droite infinie, le calcul sur la demi-droite se fait avec la précision $E/2$). Si la condition (2) n'est pas remplie, on divise ε (précision par pas) par 128 et on reprend le calcul sur la demi-droite; le calcul sur chaque Δ_i est alors plus précis. Désignons par I^2 la valeur ainsi obtenue de l'intégrale. Si

$$\min\left(\sum_{k=1}^{i+2} E_k,\ |I^2 - I^1|\right) < E\,(1 + |I^2|)$$

(E_k et i correspondent au deuxième calcul), l'intégrale est bien calculée et égale à I^2; dans le cas contraire ε est de nouveau divisé par 128 et on reprend le processus par le commencement. Si la demi-droite est exempte de singularités et si deux valeurs consécutives I_i et I_{i+1} sont nulles, le calcul s'arrête sans test des conditions (1). Passons maintenant à la description de la méthode d'intégration numérique sur le segment fini $\Delta = [a, b]$. Le calcul s'effectue selon les formules de quadrature de Markov à 7 points. Le changement de variable $x = a + (b - a)\,t$ transforme Δ en segment unité $[0, 1]$. Considérons le segment $[p, p + h] \in [0, 1]$. L'intégrale y est remplacée par la formule de quadrature correspondante et l'on calcule le terme de contrôle associé $\rho = |l(t)|$.

Si

$$\rho < \frac{\varepsilon}{b - a}, \tag{3}$$

on estime que l'intégrale sur le segment $[p, p+h]$ est déterminée. Soit $p+h<1$. Passons au segment élémentaire suivant :

$$[p_1, p_1+h_1], \qquad p_1=p+h, \qquad h_1=\min(h, 1-p_1)$$

si $\frac{\varepsilon}{128(b-a)} \leqslant \rho$, et $h_1=\min(2h, 1-p_1)$ si

$$\rho < \frac{\varepsilon}{128(b-a)}.$$

Quand on a $\frac{\varepsilon}{b-a} \leqslant \rho$, on passe au calcul de l'intégrale sur $\left[p, p+\frac{h}{2}\right]$. Il se peut que la condition (3) ne soit satisfaite pour aucun $h < H$, $H = 2^{-36}(p+2^{-24})$ (valeur voisine de zéro machine) ; ici est prévu l'arrêt de procédure à l'étiquette L avec impression préalable de l'information p et h. Le calcul commence avec $h_{\text{init}} = 1$, $p_{\text{init}} = 0$ et continue jusqu'à ce que l'intégrale ne soit calculée sur le segment [0, 1] tout entier. Si l'on se donne la singularité $[\alpha, \beta]$, pour chaque segment partiel $[p, p+h]$ on teste l'intersection des $[\alpha, \beta]$ et $[p, p+h]$. Si elle n'est pas vide et que le pas suivant $h > \tau$, h est posé égal à τ.

La procédure INT 120 pour P = 1 sert à calculer les intégrales de fonctions analytiques. Les procédures INT 110 et INT 111 lui cèdent un peu dans le cas des fonctions régulières et sont bien meilleures pour des fonctions non régulières.

Signalons certaines particularités des deux dernières procédures. Leurs corps présentent des différences insignifiantes. Pour $\varepsilon = E$ et $\alpha > \beta$, ces procédures travaillent de façon identique. La variante simplifiée (INT 110) vise à faciliter l'appel pour des séries de problèmes peu importantes. Quant à la variante compliquée INT 111, elle permet de réduire notablement la masse de calcul dans le cas de grandes séries d'intégrales et de calculer les intégrales des fonctions pourvues de singularités en pic.

La possibilité de diminuer le volume de calculs que nous venons de mentionner se réalise de manière suivante. On choisit l'intégrale type (d'ordinaire la moins apte au calcul) de la série donnée. La procédure INT 111 pose SI = 1, E égal ou inférieur à la précision voulue et $\varepsilon = \varepsilon_0$ très élevé. Les calculs nous mènent à des valeurs approchées correspondant aux coups $\varepsilon = \varepsilon_0, \varepsilon_0 2^{-k}, \ldots, \varepsilon_0 2^{-kn}$; $k = 7$ pour les intégrales du type (a), $k = 5$ pour le type (b). L'examen des résultats indique pour quel $\varepsilon = \varepsilon_0 2^{-kl}$ est réalisée la précision demandée.

Si, pour cet ε, la précision obtenue est sensiblement supérieure à celle désirée, il convient parfois d'effectuer des calculs de correction avec $\varepsilon = \varepsilon_0 2^{-1}, \ldots, \varepsilon_0 2^{-(k-1)}$. On disposera alors au total des résultats pour $\varepsilon = \varepsilon_0, \varepsilon_0 2^{-1}, \ldots$, d'où l'on choisira un ε maximal correspondant à la précision voulue.

On calcule ensuite toutes les intégrales de la série pour ε obtenu et E très grand, disons $E = 10^{19}$; un E grand garantit une seule séquence de calcul sur tout le segment.

On trouvera les intégrales multiples à l'aide de procédures ci-dessus si le compilateur admet des procédures récursives. Ainsi, on

écrit

$$I = \int_0^1 \left(\int_0^{\sqrt{1-x^2}} \exp(10(x+y))\,dy \right) dx$$

sous la forme

$$I = \int_0^1 I_1(x)\,dx$$

avec

$$I_1(x) = \int_0^{\sqrt{1-x^2}} \exp(10(x+y))\,dy,$$

et les deux fois on intègre numériquement. La fonction à intégrer de $I_1(x)$ est régulière, ce qui fait que $I_1(x)$ se calcule aisément par n'importe laquelle des procédures susmentionnées. La fonction $I_1(x)$ a une dérivée non bornée au point $x = 1$, et pour trouver $\int_0^1 I_1(x)\,dx$ on appliquera la procédure INT 120 qui convient au calcul d'intégrales des fonctions régulières.

Il paraît naturel de prime abord de calculer les intégrales sur des segments élémentaires en se servant des formules de Gauss qui, pour le même nombre de points, sont plus exactes que celles de Markov. Aussi est-ce bien ces formules qui ont été employées dans les premiers programmes standard d'intégration avec recherche automatique du pas. Cependant la pratique en a révélé un défaut dans le cas des fonctions à variation brusque. Parmi les points des formules de Gauss ne figurent pas les extrémités du segment $[p, p+h]$; si l'on calcule $\int_0^1 \exp(-Ax)\,dx$ pour A important, tous les points de la formule de Gauss appliquée au segment [0, 1] tout entier risquent de se trouver dans le domaine où la fonction à intégrer est infiniment petite; le programme s'arrête après avoir fourni une valeur approchée égale à zéro. Une situation analogue peut survenir même pour des formules contenant les extrémités (cas de $\int_0^1 x \exp(-Ax)\,dx$, A étant grand), mais le risque en est sensiblement moins grand.

Si la formule d'intégration sur un segment partiel renferme un point frontière, la valeur de la fonction en ce point sera utilisée pour trouver les intégrales sur les deux segments auxquels il appartient. C'est pourquoi le volume de calcul par segment élémentaire reste le même qu'il s'agisse de la formule de Markov à $n+1$ points ou de la formule de Gauss à n points, et les formules sont exactes pour des polynômes de même degré $2n-1$. Un autre avantage de la formule de Markov est de donner, pour son $(n+1)$-ième point, la différence divisée d'ordre n, i.e. d'évaluer la dérivée $f^{(n)}$, alors que la formule de Gauss ne permet d'évaluer que la dérivée $f^{(n-1)}$. La répartition réelle des points de la formule de Markov est visiblement plus proche de l'optimum.

4. *Procédure* INT 21. Le but de cette procédure est le calcul d'intégrales étendues à un parallélépipède rectangle à S dimensions

$$\int_{\Omega} f(X)\,dX, \quad \Omega = [a_1 \leqslant x_1 \leqslant b_1, \ \ldots, \ a_S \leqslant x_S \leqslant b_S],$$

$$X = (x_1, \ \ldots, \ x_S).$$

L'appel de procédure s'écrit:

Z: = INT 21 (S, A, F, X, EPS, P).

Paramètres formels: **entier** S (multiplicité de l'intégrale); **tableau** A [1 : 2S] (tableau donnant le domaine Ω: A [1] = a_1, A [2] = = b_1, ..., A [2S — 1] = a_S, A [2S] = b_S); **reel procedure** F (fonction procédure sans paramètres donnant la fonction à intégrer f); **tableau** X [1 : S] (argument de la fonction f); **reel** EPS (valeur de l'erreur absolue permise par pas); **reel** Z (valeur calculée de l'intégrale); **etiquette** P (étiquette à laquelle on se renvoie si la précision voulue EPS est impossible à atteindre).

DESCRIPTION DE LA MÉTHODE. Le changement de variable

$$x_i = a_i + (b_i - a_i)\,t_i, \quad i = 1, \ \ldots, \ S,$$

transforme le parallélépipède donné en cube unité

$$G = [0 \leqslant t_1 \leqslant 1, \ \ldots, \ 0 \leqslant t_S \leqslant 1],$$
$$t = (t_1, \ \ldots, \ t_S),$$

et l'intégrale est prise sur ce cube. Soit P_S le jacobien pour notre changement de variable: $P_S = (b_1 - a_1) \ \ldots \ (b_S - a_S)$. La méthode se base sur deux formules de cubature

$$\int_{c_1-h}^{c_1+h} \ldots \int_{c_S-h}^{c_S+h} g(t)\,dt \approx h^S \Sigma_q^S = Q_q^S, \quad q = 1, 2.$$

Pour trouver $\int_0^1 \ldots \int_0^1 g(t)\,dt$ on calcule Q_1^S et Q_2^S correspondant au cube G et on teste la condition

$$P_S\,|\,Q_1^S - Q_2^S\,| \leqslant \varepsilon. \tag{4}$$

Si (4) est vérifiée, on prend pour valeur approchée de l'intégrale sur G la valeur donnée par la formule Q_3^S, combinaison linéaire de Q_1^S et Q_2^S. Dans le cas contraire, on partage le cube G en 2^S cubes partiels égaux et on applique l'algorithme à chacun d'eux. Le partionnement se poursuit tant que la condition (4) n'est pas satisfaite. Si, pour h diminué de moitié, h^S devient zéro machine, le calcul s'arrête. Les formules de cubature $\sum_q^S$ avec S différents ont la forme:

1) $S=2$:

$$\Sigma_1^2 = A_1 \sum_{|i|+|j|=1} g(c_1+i\alpha h, c_2+j\alpha h) +$$

$$+A_2 \sum_{|i|+|j|=1} g(c_1+i\beta h, c_2+j\beta h) + A_3 \sum_{|i|,\,|j|=1} g(c_1+i\gamma h, c_2+j\gamma h),$$

$$\Sigma_2^2 = B_1 \sum_{|i|+|j|=1} g(c_1+i\alpha h, c_2+j\alpha h) +$$

$$+B_2 \sum_{|i|+|j|=1} g(c_1+i\beta h, c_2+j\beta h) + B_3 \sum_{|i|,\,|j|=1} g(c_1+i\gamma h, c_2+j\gamma h) +$$

$$+B_4 \sum_{|i|,\,|j|=1} g(c_1+i\nu h, c_2+j\nu h),$$

où

$\alpha = 0{,}658149897623035910,$ $\quad A_3 = 0{,}173611111111111111,$

$\beta = 0{,}549119831921783496,$ $\quad B_1 = 3{,}99942795838189963,$

$\gamma = 0{,}894427190999915878,$ $\quad B_2 = -6{,}29803906949301074,$

$\nu = 0{,}316227766016837933,$ $\quad B_3 = 0{,}124007936507936507,$

$A_1 = 1{,}0613620679054124,$ $\quad B_4 = 3{,}17460317460317460,$

$A_2 = -0{,}234973179016523356,$

$$Q_3^2 = Q_2^2.$$

La formule Q_1^2 est exacte pour tous les polynômes de degré $\leqslant 5$. La formule Q_2^2 l'est pour les polynômes de degré $\leqslant 7$.

2) $S=3$:

$$\Sigma_1^3 = \frac{8}{225}\Big(44g(c_1, c_2, c_3) +$$

$$+\frac{121}{8} \sum_{|i|,\,|j|,\,|k|=1} g\left(c_1+i\sqrt{\frac{5}{11}}h,\ c_2+j\sqrt{\frac{5}{11}}h,\ c_3+k\sqrt{\frac{5}{11}}h\right) +$$

$$+10 \sum_{|i|+|j|+|k|=1} g(c_1+ih, c_2+jh, c_3+kh)\Big),$$

$$\Sigma_2^3 = \frac{8}{1125}\Big(-\frac{1552}{5} g(c_1, c_2, c_3) +$$

$$+\frac{1573}{40} \sum_{|i|,\,|j|,\,|k|=1} g\left(c_1+i\sqrt{\frac{5}{11}}h,\ c_2+j\sqrt{\frac{5}{11}}h,\ c_3+k\sqrt{\frac{5}{11}}h\right) +$$

$$+\frac{784}{5} \sum_{|i|+|j|+|k|=1} g\left(c_1+i\sqrt{\frac{5}{14}}h,\ c_2+\right.$$

$$\left.+j\sqrt{\frac{5}{14}}h,\ c_3+k\sqrt{\frac{5}{14}}h\right)+15T_h\Big),$$

$$T_h = \sum_{|i|,\,|j|=1} g(c_1+ih,\, c_2+jh,\, c_3) + \sum_{|i|,\,|k|=1} g(c_1+ih,\, c_2,\, c_3+kh) + \sum_{|j|,\,|k|=1} g(c_1,\, c_2+jh,\, c_3+kh),$$

$$Q_3^3 = \frac{4}{9} Q_1^3 + \frac{5}{9} Q_2^3.$$

Les formules Q_1^3 et Q_2^3 sont précises pour n'importe quel polynôme de degré $\leqslant 5$. Q_3^3 l'est pour tous les polynômes de degré $\leqslant 7$.

Bibliographie du chapitre III

I [6], II [14]; § 2: I [6, 8, 16], II [14, 23]; § 3: I [16], II [14, 44], III [5]; § 5: I[13], III [27]; § 8: II [23]; §§ 9, 10: II [1], III [9, 11, 49]; § 12: I [1, 8]; § 14: III [53]; §§ 17, 18: II [1, 24], III [9, 11].

CHAPITRE IV

APPROXIMATION DES FONCTIONS ET PROBLÈMES CONNEXES

§ 1. Meilleures approximations dans un espace vectoriel normé

Considérons le problème suivant. Soit f un élément de l'espace vectoriel normé $\mathcal{R}$. On demande de trouver sa meilleure approximation par la combinaison linéaire $\sum_{i=1}^{n} c_i g_i$ des éléments linéairement indépendants $g_1, \ldots, g_n \in \mathcal{R}$ donnés. Cela équivaut à trouver un élément $\sum_{i=1}^{n} c_i^0 g_i$ tel que

$$\| f - \sum_{i=1}^{n} c_i^0 g_i \| = \Delta = \inf_{c_1, \cdots, c_n} \| f - \sum_{i=1}^{n} c_i g_i \|.$$

Si un tel élément existe, il s'appelle *meilleur approximant* (ou meilleure approximation).

THÉORÈME 1. *Le meilleur approximant existe.*

DÉMONSTRATION. Vu

$$\left| \| f - \sum_{i=1}^{n} c_i^1 g_i \| - \| f - \sum_{i=1}^{n} c_i^2 g_i \| \right| \leqslant \| \sum_{i=1}^{n} (c_i^1 - c_i^2) g_i \| \leqslant \sum_{i=1}^{n} | c_i^1 - c_i^2 | \, \| g \|,$$

la fonction

$$F_t(c_1, \ldots, c_n) = \| f - \sum_{i=1}^{n} c_i g_i \|$$

est une fonction continue en ses arguments c_j quel que soit $f \in \mathcal{R}$. Soit $\| c \|$ la norme euclidienne du vecteur $c = (c_1, \ldots, c_n)$. La fonction

$$F_0(c_1, \ldots, c_n) = \| c_1 g_1 + \ldots + c_n g_n \|$$

est continue sur la sphère unité $\| c \| = 1$ et un point $(\tilde{c}_1, \ldots, \tilde{c}_n)$ de celle-ci réalise donc sa borne inférieure $\tilde{F}$ sur la sphère. $\tilde{F} \neq 0$ puisque l'égalité $\tilde{F} = \| \tilde{c}_1 g_1 + \ldots + \tilde{c}_n g_n \| = 0$ contredit l'indépendance linéaire des éléments $g_1, \ldots, g_n$. Pour tout $c = (c_1, \ldots, c_n) \neq$

$\neq (0, \ldots, 0)$, on a l'estimation suivante :

$$\| c_1 g_1 + \ldots + c_n g_n \| = F_0(c_1, \ldots, c_n) = \\ = \| c \| F_0\left(\frac{c_1}{\| c \|}, \ldots, \frac{c_n}{\| c \|}\right) \geqslant \| c \| \widetilde{F}.$$

Soit $\gamma > \frac{2\| f \|}{\widetilde{F}}$. La fonction $F_f(c_1, \ldots, c_n)$ est continue dans la boule $\| c \| \leqslant \gamma$ et atteint donc en son point $(c_1^0, \ldots, c_n^0)$ sa borne inférieure F^0 sur la boule. On a $F^0 \leqslant F_f(0, \ldots, 0) = \| f \|$. A l'extérieur de cette boule

$$F_f(c_1, \ldots, c_n) \geqslant \| c_1 g_1 + \ldots + c_n g_n \| - \| f \| > \\ > \widetilde{F} \frac{2\| f \|}{\widetilde{F}} - \| f \| = \| f \| \geqslant F^0.$$

Ainsi,

$$F_f(c_1, \ldots, c_n) \geqslant F^0 = F_f(c_1^0, \ldots, c_n^0)$$

quels que soient $c_1, \ldots, c_n$. C.Q.F.D.

Il existe en général plusieurs meilleurs approximants.

THÉORÈME 2. *Si l'espace $\mathfrak{R}$ est strictement normé, le meilleur approximant est unique.*

DÉMONSTRATION. Raisonnons par l'absurde et supposons qu'il existe deux éléments

$$f_1 \neq f_2, \quad f_j = \sum_{i=1}^{n} a_{ij} g_i$$

tels que $\| f - f_1 \| = \| f - f_2 \| = \Delta$.

Il est clair que $\Delta \neq 0$, sinon $f_1 = f = f_2$. De plus

$$\left\| f - \frac{f_1 + f_2}{2} \right\| = \left\| \frac{f - f_1}{2} + \frac{f - f_2}{2} \right\| \leqslant \frac{\| f - f_1 \|}{2} + \frac{\| f - f_2 \|}{2} = \Delta.$$

Puisque $\frac{f_1 + f_2}{2}$ est combinaison linéaire des éléments $g_1, \ldots, g_n$, il vient $\left\| f - \frac{f_1 + f_2}{2} \right\| \geqslant \Delta$ selon la définition de Δ. Compte tenu des relations précédentes cela signifie que

$$\left\| \frac{f - f_1}{2} \right\| + \left\| \frac{f - f_2}{2} \right\| = \left\| \frac{f - f_1}{2} + \frac{f - f_2}{2} \right\|.$$

Du moment que l'espace est supposé strictement normé, on a $\frac{f - f_1}{2} = \alpha \frac{f - f_2}{2}$. Si $\alpha \neq 1$, alors $f = \frac{f_1 - \alpha f_2}{1 - \alpha}$, f est la combinaison linéaire des $g_1, \ldots, g_n$ et, partant, $\Delta = 0$. Quand $\alpha = 1$, on a $f_1 = f_2$, ce qui contredit l'hypothèse $f_1 \neq f_2$.

PROBLÈMES. 1. Démontrer qu'un espace de Hilbert est strictement normé.
2. Démontrer que l'espace $L_p(0, 1)$ avec la norme

$$\|f\|_{L_p} = \sqrt[p]{\int_0^1 |f(x)|^p \, dx}$$

est strictement normé pour $1 < p < \infty$.
3. Montrer que l'espace L_1 et l'espace $C([0, 1])$ de fonctions continues avec pour norme

$$\|f\|_C = \max_{[0, 1]} |f(x)|$$

ne sont pas strictement normés.

§ 2. Meilleure approximation dans un espace de Hilbert et sa construction

Dans le cas d'un espace de Hilbert, la recherche du meilleur approximant se ramène formellement à la résolution d'un système d'équations linéaires.

Pour obtenir ce système on procède de la façon suivante. Les coefficients a_i du meilleur approximant rendent minimale l'expression

$$\delta^2 = \Big\| \mathbf{f} - \sum_{i=1}^{n} a_i \mathbf{g}_i \Big\|^2 = \Big(\mathbf{f} - \sum_{i=1}^{n} a_i \mathbf{g}_i, \; \mathbf{f} - \sum_{i=1}^{n} a_i \mathbf{g}_i\Big).$$

En égalant à zéro les dérivées par rapport à Re a_i et Im a_i on obtient le système d'équations permettant de déterminer a_i. Comme le meilleur approximant existe, ce système admet une solution.

Formons ce système et examinons l'unicité de sa solution d'un autre point de vue.

Soit H un sous-espace vectoriel d'un espace de Hilbert $\mathcal{R}$, $\mathbf{f} \in \mathcal{R}$. Il s'agit de trouver un élément $\mathbf{h}_0 \in H$, meilleure approximation de f, i.e. tel que

$$\|\mathbf{f} - \mathbf{h}_0\| = \inf_{\mathbf{h} \in H} \|\mathbf{f} - \mathbf{h}\|.$$

LEMME 1. *Si H possède un élément $\mathbf{h}_0$, meilleur approximant de* f, *alors la différence* $\mathbf{f} - \mathbf{h}_0$ *est orthogonale à tous les éléments du sous-espace H.*

DÉMONSTRATION. Supposons qu'il existe un élément $\mathbf{h}_1 \in H$ pour lequel $(\mathbf{f} - \mathbf{h}_0, \mathbf{h}_1) = \alpha \neq 0$. Puisque l'on peut prendre $\mathbf{h}_1 / \|\mathbf{h}_1\|$ au lieu de $\mathbf{h}_1$, on admet que $\|\mathbf{h}_1\| = 1$. Considérons l'élément $\mathbf{h}_2 = \mathbf{h}_0 + \alpha \mathbf{h}_1 \in H$. On a

$$\|\mathbf{f} - \mathbf{h}_2\|^2 = (\mathbf{f} - \mathbf{h}_0 - \alpha \mathbf{h}_1, \; \mathbf{f} - \mathbf{h}_0 - \alpha \mathbf{h}_1) =$$

$$= (\mathbf{f} - \mathbf{h}_0, \; \mathbf{f} - \mathbf{h}_0) - \alpha (\mathbf{h}_1, \; \mathbf{f} - \mathbf{h}_0) - \bar{\alpha} (\mathbf{f} - \mathbf{h}_0, \; \mathbf{h}_1) + \alpha \bar{\alpha} (\mathbf{h}_1, \mathbf{h}_1),$$

mais $(\mathbf{h}_1, \mathbf{f} - \mathbf{h}_0) = \overline{(\mathbf{f} - \mathbf{h}_0, \mathbf{h}_1)} = \bar{\alpha}$. Il s'ensuit par conséquent

$$\| \mathbf{f} - \mathbf{h}_2 \|^2 = \| \mathbf{f} - \mathbf{h}_0 \|^2 - \alpha\bar{\alpha} - \bar{\alpha}\alpha + \alpha\bar{\alpha} = \\ = \| \mathbf{f} - \mathbf{h}_0 \|^2 - \alpha\bar{\alpha} < \| \mathbf{f} - \mathbf{h}_0 \|^2,$$

ce qui contredit l'hypothèse suivant laquelle $\mathbf{h}_0$ est la meilleure approximation.

LEMME 2. *Si* $(\mathbf{f} - \mathbf{h}_0, \mathbf{h}) = 0$ *pour tout* $\mathbf{h} \in H$, $\mathbf{h}_0$ *est la meilleure approximation.*

DÉMONSTRATION. On a pour $\mathbf{h} \in H$ quelconque

$$\|\mathbf{f} - \mathbf{h}\|^2 = (\mathbf{f} - \mathbf{h}_0 + \mathbf{h}_0 - \mathbf{h}, \mathbf{f} - \mathbf{h}_0 + \mathbf{h}_0 - \mathbf{h}) = \\ = (\mathbf{f} - \mathbf{h}_0, \mathbf{f} - \mathbf{h}_0) + (\mathbf{h}_0 - \mathbf{h}, \mathbf{f} - \mathbf{h}_0) + \\ + (\mathbf{f} - \mathbf{h}_0, \mathbf{h}_0 - \mathbf{h}) + (\mathbf{h}_0 - \mathbf{h}, \mathbf{h}_0 - \mathbf{h}),$$

du moment que $\mathbf{h}_0 - \mathbf{h} \in H$, le deuxième et le troisième terme s'annulent conformément à l'hypothèse. On a en définitive

$$\| \mathbf{f} - \mathbf{h} \|^2 = \| \mathbf{f} - \mathbf{h}_0 \|^2 + \| \mathbf{h} - \mathbf{h}_0 \|^2, \tag{1}$$

donc, $\| \mathbf{f} - \mathbf{h} \|^2 > \| \mathbf{f} - \mathbf{h}_0 \|^2$ pour $\mathbf{h} \neq \mathbf{h}_0$.

L'espace de Hilbert étant strictement normé, l'unicité de la meilleure approximation de la forme $a_1\mathbf{g}_1 + \ldots + a_n\mathbf{g}_n$ découle du théorème 2 du § 1. Une autre démonstration est fournie par les lemmes 1, 2 : si $\mathbf{h}_0$ est la meilleure approximation, alors $(\mathbf{f} - \mathbf{h}_0, \mathbf{h}) = 0$ pour $\mathbf{h} \in H$, et, selon (1), tout élément $\mathbf{h} \neq \mathbf{h}_0$, $\mathbf{h} \in H$, n'est pas la meilleure approximation.

Revenons au problème de la meilleure approximation de $\mathbf{f}$ par les combinaisons $\sum_{i=1}^{n} a_i\mathbf{g}_i$ des éléments $\mathbf{g}_1, \ldots, \mathbf{g}_n$ linéairement indépendants. Ces combinaisons forment un sous-espace vectoriel H. Si l'élément $\sum_{i=1}^{n} a_i\mathbf{g}_i$ est le meilleur approximant, on a, selon le lemme 1,

$$\left(\mathbf{f} - \sum_{i=1}^{n} a_i\mathbf{g}_i, \mathbf{g}_j\right) = 0 \text{ pour } j = 1, \ldots, n. \tag{2}$$

Puisque la meilleure approximation existe, il en est de même de la solution du système (2). Par ailleurs, si l'on a les relations (2), alors

$$\left(\mathbf{f} - \sum_{i=1}^{n} a_i\mathbf{g}_i, \sum_{j=1}^{n} c_j\mathbf{g}_j\right) = 0$$

quels que soient c_j, sinon $(\mathbf{f} - \sum_{i=1}^{n} a_i\mathbf{g}_i, \mathbf{h}) = 0$ pour $\mathbf{h} \in H$ quelconque ; conformément au lemme 2 cet élément $\sum_{i=1}^{n} a_i\mathbf{g}_i$ est le meilleur

approximant. Supposons que le système (2) possède une solution $(b_1, \ldots, b_n)$ autre que $(a_1, \ldots, a_n)$. Dans ce cas $g = \sum_{i=1}^{n} b_i g_i$ est lui aussi meilleur approximant et, vu son unicité,

$$\sum_{i=1}^{n} a_i g_i = \sum_{i=1}^{n} b_i g_i ;$$

l'hypothèse que $a_i \neq b_i$ pour un certain i est contraire à l'indépendance linéaire des g_i.

Ainsi, nous avons montré que le système (2) admet toujours une solution et une seule. On met ce système sous la forme

$$\sum_{i=1}^{n} a_i (g_i, g_j) = (f, g_j), \quad j = 1, \ldots, n. \tag{3}$$

On trouve donc les coefficients associés au meilleur approximant en résolvant le système non dégénéré (3).

La matrice

$$G(g_1, \ldots, g_n) = [(g_i, g_j)]$$

est dite *matrice de Gram* de ce système.

En étudiant les propriétés de la matrice de Gram on utilisera l'égalité formelle

$$G(g_1, \ldots, g_n) = \begin{pmatrix} g_1 \\ \cdots \\ g_n \end{pmatrix} (g_1, \ldots, g_n)$$

PROBLÈME. Démontrer que $G(g_1, \ldots, g_n) \geqslant 0$ et qu'on a l'égalité si et seulement si $g_1, \ldots, g_n$ sont linéairement dépendants. Démontrer l'inégalité

$$G(g_1, \ldots, g_{n+1}) \leqslant G(g_1, \ldots, g_n)(g_{n+1}, g_{n+1}).$$

Soit g la meilleure approximation de f. Alors, selon le lemme 1, on a $(f - g, h) = 0$ pour $h \in H$. Portant $h_0 = g$ et $h = 0$ dans (1), il vient

$$\| f \|^2 = \| f - g \|^2 + \| g \|^2.$$

Si les éléments g_i forment un système orthonormé $(g_i, g_j) = \delta_{i-j}$, le système (3) se récrit

$$a_j = (f, g_j) ; \tag{4}$$

on a le meilleur approximant:

$$g = \sum_{i=1}^{n} (f, g_i) g_i,$$

et l'écriture commode de la quantité $\|f-g\|^2$:

$$\|f-g\|^2=\|f\|^2-\|g\|^2=(f, f)-\Big(\sum_{i=1}^{n} a_i g_i, \sum_{i=1}^{n} a_i g_i\Big)=$$

$$=(f, f)-\sum_{i, j=1}^{n} a_i \bar{a}_j \delta_{i-j}=(f, f)-\sum_{i=1}^{n}|a_i|^2=(f, f)-\sum_{i=1}^{n}|(f, g_i)|^2.$$

Puisque $\|f-g\|^2 \geqslant 0$, l'égalité

$$\|f-g\|^2=(f, f)-\sum_{i=1}^{n}|(f, g_i)|^2$$

implique en particulier l'inégalité bien connue de Bessel:

$$(f, f) \geqslant \sum_{i=1}^{n}|(f, g_i)|^2.$$

Si les éléments initiaux ne forment pas un système orthonormé, on les orthogonalise à l'aide de l'algorithme du chapitre II.

Pour résoudre des problèmes assez compliqués on a souvent besoin de calcul d'un jeu de coefficients $a_j=(f, g_j)$, $j=1, \ldots, M$ (M relativement grand).

Etudions les procédés de calcul des a_j sur l'exemple des intégrales

$$a_j=(f, Q_j)=\int_a^b f(x) Q_j(x) p(x)\, dx, \qquad p(x)>0,$$

où $Q_j(x)$ réels constituent un système orthonormé dans un espace muni du produit scalaire

$$(f, g)=\int_a^b f(x) g(x) p(x)\, dx.$$

Les intégrales a_j se calculent d'une façon approchée par la formule de Gauss (§ 3, ch. II) correspondant à la fonction de poids donnée

$$a_j=\int_a^b f(x) Q_j(x) p(x)\, dx \approx \sum_{q=1}^{N} D_q^N f(x_q^N) Q_j(x_q^N). \tag{5}$$

Si f est un polynôme de degré l, alors le degré de $f(x) Q_j(x)$ est $l+j$. Puisque la formule de Gauss est exacte pour les polynômes de degré $2N-1$, la formule (5) est stricte si $f(x)$ est un polynôme de degré $2N-j-1$.

La résolution d'un problème exigeait qu'on calcule plusieurs milliers d'intégrales du type

$$a_j=\int_a^b f(x) Q_j(x) p(x)\, dx, \qquad j=0, \ldots, M.$$

Aucune tentative d'utiliser les programmes d'intégration avec recherche automatique du pas ou les formules (5) n'a abouti. En analysant le problème on a constaté que sa résolution par calcul successif des a_j exigeait un temps machine prohibitif.

Le point le plus délicat était le calcul très pénible des valeurs de $f(x)$. On a donc adopté la méthode suivante. Les intégrales a_j ont été divisées en groupes

$$(a_0, \ldots, a_{n_1}), (a_{n_1+1}, \ldots, a_{n_2}), \ldots, (a_{n_{k-1}+1}, \ldots, a_{n_k}), \quad n_k = M.$$

Pour chaque groupe on a déterminé un nombre N_l suffisant pour que la formule (5) assure au calcul de chacun de ses éléments la précision demandée. Ensuite, on a calculé des valeurs approchées pour toutes les intégrales du groupe à la fois:

$$a_j^{(N_l)} = \sum_{q=1}^{N_l} D_q^{N_l} f(x_q^{N_l}) Q_j(x_q^{N_l}).$$

Chaque $f(x_q^{N_l})$ était calculée une seule fois pour le groupe tout entier. On a également réalisé un gain de temps machine supplémentaire provenant du calcul simultané des valeurs de tous les polynômes $Q_j(x_q^{N_l})$ en un seul point par des relations de récurrence. En répartissant les coefficients a_j on a tenu compte de la masse d'opérations qu'exigeait le calcul des valeurs de la fonction $f(x)$.

Dans un autre cas semblable il s'est avéré plus rationnel, du point de vue de l'économie de temps machine, de s'abstenir d'une pareille division, de prendre un N correspondant à a_M et de calculer avec toutes les intégrales à la fois. Le volume de calcul a été réduit d'environ 100 fois par rapport au procédé initial.

Nous avons déjà parlé de l'habitude des clients d'apprécier l'efficacité d'une méthode d'après la part de travail utile. C'est d'ailleurs la raison pour laquelle ils se sont opposés un certain temps à la procédure décrite en alléguant la « trop grande précision » des calculs pour j petits, qui entraînait beaucoup de travail à vide. Les calculateurs les ont convaincus alors de la haute efficacité de la méthode en arguant du pourcentage de travail utile: le volume de calcul de toute la série d'intégrales n'est supérieur que de 20 % à celui de l'intégrale a_M par la formule (5), ce qui fait que la part de travail utile constitue quelque 80 pour cent. L'argument a porté, et les clients ont accepté la méthode. Si la fonction $f(x)$ est mal approchée par des polynômes, on procède comme suit: on calcule toutes les intégrales simultanément à l'aide de l'algorithme d'intégration, la recherche du pas étant automatique et les points d'intégration restant identiques pour toutes les intégrales, les mêmes que pour l'intégrale a_M qui est « la pire » de toutes. Pour réaliser cet algorithme la procédure standard d'intégration avec recherche automatique du pas ne convient plus, et un nouveau programme s'impose. Il existe une autre possibilité, à savoir ne pas calculer directement toutes les valeurs de $f(x)$, mais déterminer certaines d'entre elles en interpolant sur des valeurs déjà calculées.

Considérons un problème d'un type différent dont la résolution recourt dans un certain sens aux constructions de ce paragraphe. On demande de construire un polynôme d'interpolation de degré $N-1$ basé sur les abscisses $x_1^N, \ldots, x_N^N$, zéros du polynôme $Q_N(x)$ de degré N d'un système orthonormé $\{Q_N(x)\}$ relativement au poids $p(x)$. (Dans le chapitre II nous avons parlé de l'interpolation sur les zéros des polynômes de Tchebycheff.) Cherchons ce

polynôme sous forme de combinaison linéaire

$$P_{N-1}(x) = \sum_{j=0}^{N-1} a_j Q_j(x).$$

Pour m, $n < N$, le degré du polynôme $Q_m(x)\, Q_n(x)$ est au plus $2N - 2$. Aussi la formule de quadrature de Gauss à N points est-elle exacte pour ce polynôme

$$\sum_{q=1}^{N} D_q^N Q_m(x_q^N)\, Q_n(x_q^N) = \int_a^b Q_m(x)\, Q_n(x)\, p(x)\, dx = \delta_{n-m}. \qquad (6)$$

Les vecteurs

$$\mathbf{Q}_m = \{Q_m(x_1^N), \ldots, Q_m(x_N^N)\}$$

forment donc pour $m < N$ un système orthonormé relativement au produit scalaire

$$(\mathbf{y}, \mathbf{z}) = \sum_{q=1}^{N} D_q^N y_q z_q ; \qquad (7)$$

et le vecteur $\mathbf{f} = (f(x_1^N), \ldots, f(x_N^N))$ se décompose suivant ce système de vecteurs :

$$\mathbf{f} = \sum_{j=0}^{N-1} d_j Q_j$$

avec $d_j = (\mathbf{f}, \mathbf{Q}_j)$.

Le polynôme

$$P_{N-1}(x) = \sum_{j=0}^{N-1} d_j Q_j(x)$$

est bien le polynôme cherché.

Dans les calculs, on se sert surtout des fonctions de poids $p(x) = \dfrac{1}{\sqrt{1-x^2}}$ sur $[-1, 1]$ et $p(x) = 1$. Pour la première on connaît explicitement les polynômes orthogonaux et la formule de quadrature de Gauss (il s'agit des polynômes de Tchebycheff et de la formule d'Hermite), et pour la seconde on connaît les polynômes et on dispose de tableaux suffisamment fournis des points d'interpolation et des poids des formules de quadrature de Gauss.

§ 3. Transformation de Fourier discrète

Supposons que $f(x)$ périodique de période 1 se développe en série de Fourier

$$f(x) = \sum_{q=-\infty}^{\infty} a_q \exp(2\pi i q x) \qquad (1)$$

avec

$$\sum_{q=-\infty}^{\infty} |a_q| < \infty. \qquad (2)$$

Prenons les valeurs de cette fonction sur un réseau de points $x_l = l/N$, où l, N (fixe) sont entiers, et notons $f(x_l) = f_l$. Soit $q_2 - q_1 = kN$, k entier; alors $q_2 x_l - q_1 x_l = kNx_l = kl$ (kl entier). Donc,

$$\exp(2\pi i q_1 x) = \exp(2\pi i q_2 x) \tag{3}$$

aux nœuds du réseau. C'est pourquoi, si $f(x)$ n'est considérée que sur le réseau de points x_l, on peut réduire les termes semblables dans (1):

$$f_l = \sum_{q=0}^{N-1} A_q \exp(2\pi i q x_l) \tag{4}$$

où

$$A_q = \sum_{s=-\infty}^{\infty} a_{q+sN}.$$

Si l'on se donne dès le début une fonction définie seulement sur un réseau, on peut la représenter sur ce réseau par (1). En effet, une telle fonction est prolongeable sur toute la droite si on trouve ses valeurs entre les nœuds par interpolation linéaire. Une fonction continue dérivable par morceaux vérifie (2), et après avoir réduit les termes semblables sur le réseau on aboutit à (4). Définissons le produit scalaire des fonctions discrètes:

$$(f, g) = \frac{1}{N} \sum_{l=0}^{N-1} f_l \bar{g}_l.$$

Le facteur $1/N$ a été introduit pour concorder les relations obtenues avec le cas continu: si $f(x)$ et $g(x)$ sont continues sur $[0, 1]$, alors

$$(f, g) \to \int_0^1 f(x)\, \bar{g}(x)\, dx$$

pour $N \to \infty$.

Les fonctions $g_q(x_l) = \exp(2\pi i q x_l)$ forment pour $0 \leqslant q < N$ un système orthonormé relativement au produit scalaire ainsi défini. En effet,

$$(g_q, g_j) = \frac{1}{N} \sum_{l=0}^{N-1} \exp\left(2\pi i \frac{q-j}{N} l\right).$$

Pour $q \neq j$ on obtient en sommant la progression géométrique

$$(g_q, g_j) = \frac{1}{N} \frac{\exp(2\pi i (q-j)) - 1}{\exp\left(2\pi i \frac{q-j}{N}\right) - 1} = 0.$$

Puisque $(g_q, g_q) = 1$, on a finalement

$$(g_q, g_j) = \delta_{q-j} \text{ pour } 0 \leqslant q,\ j < N. \tag{5}$$

Multipliant scalairement (4) par g_j, il vient l'égalité

$$A_j = (f, g_j) = \frac{1}{N} \sum_{l=0}^{N-1} f_l \exp(-2\pi i j x_l); \tag{6}$$

évidemment

$$A_j \to a_j = \int_0^1 f(x) \exp(-2\pi i j x)\, dx$$

pour $N \to \infty$ et j fixe.

Cela ne veut nullement dire que

$$f(x) \approx \sum_{j=0}^{N-1} A_j \exp(2\pi i j x).$$

Pourquoi ? Supposons, pour plus de simplicité, que

$$f(x) = \sum_{-k}^{k} a_j \exp(2\pi i j x),$$

où $2k < N$. L'égalité (4) donne la représentation des coefficients A_j :

$$A_j = \begin{cases} a_j & \text{pour } 0 \leqslant j \leqslant k, \\ 0 & \text{pour } k < j < N-k, \\ a_{j-N} & \text{pour } N-k \leqslant j < N. \end{cases}$$

Le second membre de cette égalité approchée inexacte est donc

$$\sum_{j=0}^{k} a_j \exp(2\pi i j x) + \exp(2\pi i N x) \sum_{j=-k}^{-1} a_j \exp(2\pi i j x).$$

Il coïncide avec $f(x)$ en x_l et en est éloigné en dehors de ces points.

En substituant la variable de sommation $s' = s + k$ dans la définition de A_q on se persuade que $A_{q_1} = A_{q_2}$ si $q_1 - q_2 = kN$ avec k entier. Compte tenu de (3) on récrit (4)

$$f_l = \sum_{-N/2 < q \leqslant N/2} A_q \exp(2\pi i q x_l). \tag{7}$$

Si $f(x)$ est suffisamment régulière, les quantités $|a_j|$ décroissent rapidement avec l'augmentation de j, donc $A_q \approx a_q$ pour q petits et A_q et a_q sont faibles dès que q sont grands. En résumé on a des

raisons d'espérer que l'égalité approchée

$$f(x) \approx \sum_{-N/2<q\leqslant N/2} A_q \exp(2\pi i q x) \tag{8}$$

sera valable en tous les x.

Rappelons que (8) devient une égalité aux points du réseau. Le procédé d'approximation d'une fonction par le second membre de (8) est l'*interpolation trigonométrique*. La relation (7) est dite *série de Fourier finie* ou *discrète* et A_q *coefficients de Fourier discrets*.

Le fait de négliger l'égalité établie des fonctions

$$\exp(2\pi i q_1 x) \text{ et } \exp(2\pi i q_2 x)$$

aux nœuds du réseau pour $q_1 - q_2 = kN$ est souvent la cause des relations fausses.

Dans un problème d'ingénieur, on demandait de calculer la première fréquence propre de vibrations d'une construction. On a décidé d'écrire une équation non stationnaire décrivant le processus vibratoire et d'imprimer la courbe représentative dont l'examen devrait fournir la fréquence. On a résolu l'équa-

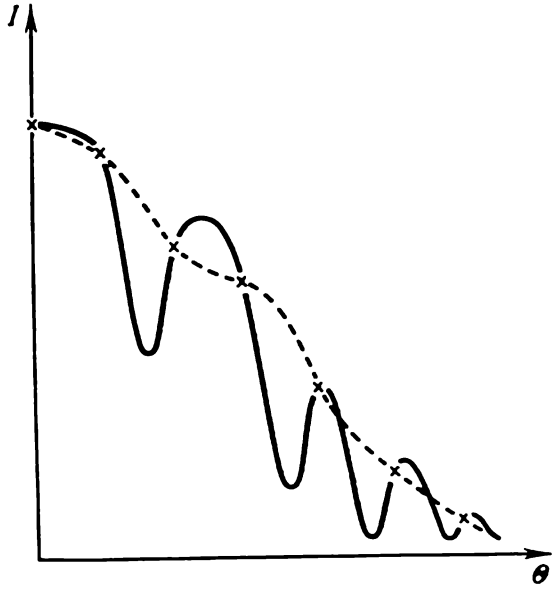


Fig. 4.3.1

tion correspondante, que nous écrivons conventionnellement $x'' = F(x)$, par la méthode des différences finies. On a testé le résultat en refaisant les calculs avec un pas diminué de moitié. Les courbes ainsi obtenues ont coïncidé à 10 % près. Cependant, on a constaté dans l'expérience que la fréquence calculée s'écartait de la fréquence vraie de plusieurs dizaines de fois. La cause en était le pas $1/N$ adopté pour construire le graphique de la solution, qui était sensiblement supérieur à la période de vibration. La solution était voisine de la fonction

const·exp $(2\pi iqt)$ avec q/N proche de $2k$ pair. Aussi le réseau de pas $1/N$ tout comme le réseau deux fois plus fin (avec le pas $1/(2N)$) ont-ils donné tous deux le graphique de la même fonction const·exp $(2\pi i\,(q-2kN)\,t)$.

Une autre fois on a obtenu une solution absurde en calculant le diagramme de directivité d'une antenne. On a essayé en vain de détecter l'erreur dans le programme, la méthode de résolution ou la description physique du problème. Or, la cause d'erreur était la même que dans l'exemple précédent: le graphique d'une fonction fortement oscillante était imprimé sur un réseau très espacé. La fig. 4.3.1. visualise en trait continu le graphique réel de la section du diagramme et en pointillé celui obtenu par interpolation des x calculés.

Il existe une correspondance entre le problème d'approximation des fonctions par des combinaisons linéaires de polynômes de Tchebycheff et leur approximation par des polynômes trigonométriques. Supposons que la fonction $f(x)$ est approchée sur le segment $[-1, 1]$ par les combinaisons linéaires $\sum_{j=0}^{m-1} a_j T_j(x)$. Par le changement de variables $x = \cos t$ le problème se ramène à l'approximation de la fonction $f(\cos t)$ par la combinaison linéaire

$$\sum_{j=0}^{m-1} a_j T_j(\cos t) = \sum_{j=0}^{m-1} a_j \cos(jt).$$

On a l'égalité

$$(f, g)_1 = \int_{-1}^{1} \frac{f(x)\,\bar{g}(x)}{\sqrt{1-x^2}}\,dx = \int_0^{\pi} f(\cos\theta)\,\bar{g}(\cos\theta)\,d\theta = (f, g)_2.$$

Par conséquent, chercher la meilleure approximation de $f(x)$ pour la norme associée au produit scalaire $(f, g)_1$ revient à chercher l'approximation de $f(\cos\theta)$ pour la norme associée au produit scalaire $(f, g)_2$. On note la même correspondance dans le cas de l'interpolation et de la meilleure approximation dans une métrique uniforme. L'interpolation de la fonction par un polynôme basé sur les points $x_j = \cos\left(\pi\,\frac{2j-1}{2m}\right)$, zéros du polynôme de Tchebycheff $T_m(x)$, se ramène alors à l'interpolation de $f(\cos\theta)$ par le polynôme trigonométrique $\sum_{j=0}^{m-1} a_j \cos(jt)$ avec les points $t_j = \pi\,\frac{2j-1}{2N}$ formant un réseau régulier.

§ 4. Transformation de Fourier rapide

La transformation de Fourier discrète et son inverse

$$(f_0, \ldots, f_{N-1}) \Longleftrightarrow (A_0, \ldots, A_{N-1})$$

figurent dans de nombreux problèmes. En les réalisant directement par les formules (3.4) et (3.6) on effectue $O(N^2)$ opérations arithmétiques. Voyons s'il est possible de réduire ce chiffre et proposons-nous, pour fixer les idées, de calculer les coefficients A_q pour les

valeurs données de la fonction. Pour construire les algorithmes de *transformation de Fourier rapide* on part du fait que, pour N composé, on peut isoler dans les termes à droite de (3.6) des groupes figurant dans les expressions des divers coefficients A_q.

Posons d'abord $N = p_1 p_2$, $p_1, p_2 \neq 1$. Mettons q, j compris dans les limites $0 \leqslant q, j < N$ sous la forme $q = q_1 + p_1 q_2$, $j = j_2 + p_2 j_1$, avec $0 \leqslant q_1, j_1 < p_1$, $0 \leqslant q_2, j_2 < p_2$. On aboutit à la chaîne de relations

$$A_q = A(q_1, q_2) = \frac{1}{N} \sum_{j=0}^{N-1} f_j \exp\left(-2\pi i \frac{qj}{N}\right) =$$

$$= \frac{1}{N} \sum_{j_1=0}^{p_1-1} \sum_{j_2=0}^{p_2-1} f_{j_2+p_2 j_1} \exp\left(-2\pi i \frac{(q_1+p_1q_2)(j_2+p_2j_1)}{p_1p_2}\right).$$

Moyennant l'égalité

$$\frac{(q_1+p_1q_2)(j_2+p_2j_1)}{p_1p_2} = q_2 j_1 + \frac{qj_2}{N} + \frac{j_1 q_1}{p_1}$$

et la relation précédente, on trouve

$$A(q_1, q_2) = \frac{1}{p_2} \sum_{j_2=0}^{p_2-1} A^{(1)}(q_1, j_2) \exp\left(-2\pi i \frac{qj_2}{N}\right)$$

avec

$$A^{(1)}(q_1, j_2) = \frac{1}{p_1} \sum_{j_1=0}^{p_1-1} f_{j_2+p_2 j_1} \exp\left(-2\pi i \frac{j_1 q_1}{p_1}\right).$$

Le calcul direct de tous les $A^{(1)}(q_1, j_2)$ exige $O(p_1^2 p_2)$ opérations arithmétiques, et le calcul suivant de $A(q_1, q_2)$ encore $O(p_1 p_2^2)$. Par conséquent, pour $p_1, p_2 = O(\sqrt{N})$, on doit effectuer en tout $O(N^{3/2})$ opérations. On construit de façon analogue, avec $N = p_1 \ldots p_r$, l'algorithme de calcul de l'ensemble des valeurs de A_q, pour lequel les opérations sont au nombre de $CN(p_1 + \ldots + p_r)$ au plus (C une constante absolue). Ecrivons les formules de calcul correspondantes pour le cas le plus courant $p_1 = \ldots = p_r = 2$. Mettons q, j sous la forme

$$q = \sum_{k=1}^{r} q_k 2^{k-1}, \quad j = \sum_{m=1}^{r} j_{r+1-m} 2^{m-1},$$

où $q_k, j_m = 0, 1$. La quantité $qj2^{-r}$ s'écrit

$$qj2^{-r} = \sum_{m=1}^{r} qj_{r+1-m} 2^{m-1-r} = \sum_{m=1}^{r} \left(\sum_{k=1}^{r} q_k 2^{k+m-r-2}\right) j_{r+1-m} =$$

$$= \sum_{m=1}^{r} \left(\sum_{k=1}^{r-m+1} q_k 2^{k+m-r-2}\right) j_{r+1-m} + s$$

avec s entier, somme de tous les termes

$$q_k j_{r+1-m} 2^{k+m-r-2},$$

pour lesquels $k+m-r-2 \geqslant 0$. Il est clair que

$$\exp\left(-2\pi i \frac{qj}{N}\right) = \exp\left(-2\pi i \left(\frac{qj}{N} - s\right)\right).$$

Donc

$$A(q_1, \ldots, q_r) = A_q = \frac{1}{N} \sum_{j=0}^{N-1} f_j \exp\left(-2\pi i \frac{jq}{N}\right) =$$

$$= \frac{1}{2} \sum_{j_r=0}^{1} \cdots \frac{1}{2} \sum_{j_1=0}^{1} f_{j_r+2j_{r-1}+\ldots+2^{r-1}j_1} \times$$

$$\times \exp\left(-2\pi i \sum_{m=1}^{r} \left(\sum_{k=1}^{r-m+1} q_k 2^{k+m-r-2}\right) j_{r+1-m}\right).$$

On obtient en regroupant les termes:

$$A(q_1, \ldots, q_r) = \frac{1}{2} \sum_{j_r=0}^{1} \exp\left(-2\pi i j_r 2^{-r} \sum_{k=1}^{r} q_k 2^{k-1}\right) \times$$

$$\times \left(\frac{1}{2} \sum_{j_{r-1}=0}^{1} \exp\left(-2\pi i j_{r-1} 2^{1-r} \sum_{k=1}^{r-1} q_k 2^{k-1}\right) \times \ldots\right.$$

$$\left. \ldots \times \left(\frac{1}{2} \sum_{j_1=0}^{1} \exp\left(-2\pi i j_1 2^{-1} q_1\right) f_{j_r+2j_{r-1}+\ldots+2^{r-1}j_1}\right) \ldots\right).$$

On écrit cette relation sous forme d'une suite de relations de récurrence

$$A^{(m)}(q_1, \ldots, q_m; j_{m+1}, \ldots, j_r) =$$

$$= \frac{1}{2} \sum_{j_m=0}^{1} \exp\left(-2\pi i j_m 2^{-m} \sum_{k=1}^{m} q_k 2^{k-1}\right) \times$$

$$\times A^{(m-1)}(q_1, \ldots, q_{m-1}; j_m, \ldots, j_r), \quad m = 1, \ldots, r,$$

où

$$A^{(0)}(j_1, \ldots, j_r) = f_{j_r+j_{r-1}2+\ldots+j_1 2^{r-1}},$$

$$A^{(r)}(q_1, \ldots, q_r) = A(q_1, \ldots, q_r).$$

On passe de l'ensemble $A^{(m-1)}$ à l'ensemble $A^{(m)}$ moyennant $O(N)$ opérations mathématiques et logiques; vu qu'il y a r pas, le nombre

total d'opérations est

$$O(Nr) = O(N \log_2 N).$$

En se servant de cet ensemble de relations on commet une erreur d'accumulation moindre par comparaison avec les formules (3.6). Ce procédé offre également un certain avantage quant au calcul des exponentielles faisant partie des formules. Pour obtenir $A^{(m)}$ on utilise les valeurs $\exp(-2\pi ij2^{-m})$, $j = 0, 1, \ldots, 2^m - 1$; en particulier, ce sont 1 et -1 lorsque $m = 1$. Pour trouver $A^{(m+1)}$ on a encore besoin des valeurs $\exp(-2\pi ij2^{-m-1})$ pour j impairs vérifiant l'inégalité $0 \leqslant j < 2^{m+1}$; ces dernières valeurs se calculent à l'aide de quantités connues, notamment par le recours aux relations ($m \geqslant 2$):

$$\exp(-2\pi ij2^{-m-1}) = \frac{\exp\left(-2\pi i \frac{j+1}{2} 2^{-m}\right) + \exp\left(-2\pi i \frac{j-1}{2} 2^{-m}\right)}{2\cos(\pi 2^{-m})},$$

où

$$\cos(\pi 2^{-m}) = \sqrt{\frac{1+\cos(\pi 2^{1-m})}{2}}$$

pour $m \geqslant 1$.

Dans certains cas on arrive à réduire le nombre d'opérations de plusieurs fois encore. Nous en avons déjà parlé: on s'est donné une fonction réelle $f(t) = g(\cos t)$ connue aux points $t_l = \frac{\pi(2l-1)}{2N}$; on demandait de trouver les coefficients du polynôme d'interpolation

$$\sum_{j=0}^{N-1} A_j \cos jt.$$

Voici un autre cas. Connaissant les valeurs de la fonction

$$\sum_{j=1}^{N/2-1} A_j \sin 2\pi jx$$

aux points $x_l = l/N$, $0 < l < N/2$ (N pair), trouver les coefficients A_j.

PROBLÈME. Trouver les coefficients c_j du produit des deux polynômes

$$\left(\sum_{j=0}^{N-1} a_j x^j\right)\left(\sum_{j=0}^{N-1} b_j x^j\right) = \sum_{j=0}^{2N-2} c_j x^j.$$

Montrer que $O(N \log_2 N)$ opérations suffisent bien.

§ 5. Meilleure approximation uniforme

Si dans un espace vectoriel normé la norme n'est pas déterminée à l'aide du produit scalaire, la recherche du meilleur approximant se complique sensiblement. Considérons un problème type qu'on rencontre, en particulier, lors de la rédaction de programmes standard de calcul de fonctions.

Soit $\mathscr{R}$ un espace de fonctions réelles bornées définies sur le segment $[a, b]$ de l'axe réel avec pour norme

$$\| f \| = \sup_{[a, b]} | f(x) |.$$

On cherche la meilleure approximation de la forme

$$Q_n(x) = \sum_{j=0}^{n} a_j x^j.$$

Conformément au théorème 1 du § 1 il existe la meilleure approximation, c'est-à-dire un polynôme $Q_n^0(x)$ tel que

$$E_n(f) = \| f - Q_n^0 \| \leqslant \| f - Q_n \|$$

pour n'importe quel $Q_n(x)$ de degré n. Ce polynôme $Q_n^0(x)$ s'appelle *polynôme de meilleure approximation uniforme*. Dans la suite on établira les conditions nécessaires et suffisantes pour qu'un polynôme soit la meilleure approximation pour une fonction continue.

THÉORÈME (DE VALLÉE-POUSSIN). *S'il existe* $n + 2$ *points* $x_0 < \ldots < x_{n+1}$ *du segment* $[a, b]$ *tels que*

$$\operatorname{sg}((f(x_i) - Q_n(x_i))(-1)^i) = \text{const},$$

i.e. *en passant d'un point* x_i *au point suivant* x_{i+1}, *la quantité* $f(x) - Q_n(x)$ *change de signe, alors*

$$E_n(f) \geqslant \mu = \min_{i=0, \ldots, n+1} | f(x_i) - Q_n(x_i) |. \tag{1}$$

DÉMONSTRATION. La proposition est évidente pour $\mu = 0$. Soit $\mu > 0$. Supposons le contraire, c'est-à-dire que pour le polynôme de meilleure approximation $Q_n^0(x)$

$$\| Q_n^0 - f \| = E_n(f) < \mu.$$

On a

$$\operatorname{sg}(Q_n(x) - Q_n^0(x)) = \operatorname{sg}((Q_n(x) - f(x)) - (Q_n^0(x) - f(x))).$$

Aux points x_i le premier terme est supérieur en module au deuxième, aussi $\operatorname{sg}(Q_n(x_i) - Q_n^0(x_i)) = \operatorname{sg}(Q_n(x_i) - f(x_i))$. Par conséquent, le polynôme $Q_n(x) - Q_n^0(x)$ de degré n change de signe $n + 1$ fois. Nous avons abouti à une contradiction.

THÉORÈME D'ALTERNANCE DE TCHEBYCHEFF. *Pour qu'un polynôme* $Q_n(x)$ *soit le polynôme de meilleure approximation de la fonction continue* $f(x)$, *il faut et il suffit que le segment* $[a, b]$ *contienne au moins* $m = n + 2$ *points* $x_0 < \ldots < x_{n+1}$ *tels que*

$$f(x_i) - Q_n(x_i) = \alpha (-1)^i \| f - Q_n \|.$$

$i = 0, \ldots, n+1$, $\alpha = 1$ (*ou* $\alpha = -1$) *simultanément pour tous les* i.

On dit des points $x_0, \ldots, x_{n+1}$ satisfaisant aux conditions du théorème qu'ils forment le *support de Tchebycheff*.

DÉMONSTRATION. SUFFISANCE. Notons L la quantité $\| f - Q_n \|$. En appliquant (1), il vient

$$L = \mu \leqslant E_n(f), \text{ mais } E_n(f) \leqslant \| f - Q_n \| = L$$

selon la définition de $E_n(f)$. Aussi $E_n(f) = L$ et le polynôme donné est bien meilleure approximation uniforme.

NÉCESSITÉ. Supposons que notre polynôme $Q_n(x)$ soit meilleure approximation uniforme. Désignons par y_1 la borne inférieure des points $x \in [a, b]$ en lesquels $|f(x) - Q_n(x)| = L$; l'existence de ce point découle de la définition de L. Vu la continuité de $f(x) - Q_n(x)$, on a $|f(y_1) - Q_n(y_1)| = L$. Pour fixer les idées, on se bornera au cas $f(y_1) - Q_n(y_1) = +L$. Désignons par y_2 la borne inférieure de tous les points $x \in (y_1, b]$ en lesquels $f(x) - Q_n(x) = -L$, par y_{k+1} celle des points $x \in (y_k, b]$ en lesquels $f(x) - Q_n(x) = (-1)^k L$, ... On a pour n'importe quel k, vu la continuité de $f(x) - Q_n(x)$:

$$f(y_{k+1}) - Q_n(y_{k+1}) = (-1)^k L.$$

Poursuivons tant qu'on a $y_m = b$ ou un y_m tel que $|f(x) - Q_n(x)| < L$ pour $y_m < x \leqslant b$. Si $m \geqslant n+2$, le théorème a lieu. Supposons le contraire.

Du moment que $f(x) - Q_n(x)$ est continu, on peut indiquer, quel que soit k $(1 < k \leqslant m)$, un point z_{k-1} tel que $|f(x) - Q_n(x)| < L$ avec $z_{k-1} \leqslant x < y_k$; posons $z_0 = a$, $z_m = b$. Par construction les segments $[z_{i-1}, z_i]$, $i = 1, \ldots, m$, possèdent des points (tels que y_i) en lesquels $f(x) - Q_n(x) = (-1)^{i-1} L$ et ne présentent pas de points où $f(x) - Q_n(x) = (-1)^i L$. Posons

$$v(x) = \prod_{j=1}^{m-1} (z_j - x), \quad Q_n^d(x) = Q_n(x) + dv(x), \quad d > 0,$$

et examinons l'allure de la différence

$$f(x) - Q_n^d(x) = f(x) - Q_n(x) - dv(x)$$

sur les segments $[z_{j-1}, z_j]$. Choisissons par exemple le segment $[z_0, z_1]$. Sur $[z_0, z_1)$ on a $v(x) > 0$, donc

$$f(x) - Q_n^d(x) \leqslant L - dv(x) < L.$$

D'autre part, $f(x) - Q_n(x) > -L$ sur $[z_0, z_1]$, ce qui fait que pour d suffisamment petits, par exemple

$$d < d_1 = \frac{\min\limits_{[z_0, z_1]} |f(x) - Q_n(x) + L|}{\max\limits_{[z_0, z_1]} |v(x)|},$$

on a sur $[z_0, z_1)$:

$$f(x) - Q_n^d(x) > -L.$$

De plus,

$$|f(z_1) - Q_n^d(z_1)| = |f(z_1) - Q_n(z_1)| < L.$$

Ainsi

$$|f(x) - Q_n^d(x)| < L$$

sur ce segment pour un d assez petit. En raisonnant de façon analogue sur les autres segments, nous pouvons indiquer un d_0 petit tel que sur tous les segments on ait

$$|f(x) - Q_n^{d_0}(x)| < L;$$

$v(x)$ et, partant, $Q_n^{d_0}(x)$ sont des polynômes de degré n.

Cela est contraire à notre hypothèse selon laquelle $Q_n(x)$ est le polynôme de meilleure approximation et $m < n + 2$. Le théorème est démontré.

Cette démonstration du théorème de Tchebycheff, ainsi que la méthode itérative du § 7 qui va nous permettre d'obtenir le polynôme de meilleure approximation uniforme, constitue une sorte d'impasse dans l'édifice de notre cours. En effet, ni les méthodes ni les résultats ne trouvent aucune application dans l'exposé suivant. N'empêche qu'ils peuvent intéresser le lecteur en tant qu'exemple typique d'étude des problèmes de la programmation mathématique, une branche récente des mathématiques appliquées qui se développe à pas de géant.

THÉORÈME D'UNICITÉ. *Il existe un seul polynôme, meilleure approximation uniforme de la fonction continue.*

DÉMONSTRATION. Supposons qu'il en existe deux de degré n ;

$$Q_n^1(x) \neq Q_n^2(x), \quad \|f - Q_n^1\| = \|f - Q_n^2\| = E_n(f).$$

Alors

$$\left\|f - \frac{Q_n^1 + Q_n^2}{2}\right\| \leqslant \left\|\frac{f - Q_n^1}{2}\right\| + \left\|\frac{f - Q_n^2}{2}\right\| = E_n(f).$$

Le polynôme $\dfrac{Q_n^1(x) + Q_n^2(x)}{2}$ est donc lui aussi la meilleure approximation uniforme. Soient $x_0, \ldots, x_{n+1}$ les points correspondants

du support de Tchebycheff. Alors

$$\left|\frac{Q_n^1(x_i)+Q_n^2(x_i)}{2}-f(x_i)\right|=E_n(f), \quad i=0, \ldots, n+1,$$

ou

$$|(Q_n^1(x_i)-f(x_i))+(Q_n^2(x_i)-f(x_i))|=2E_n(f).$$

Puisque $|Q_n^k(x_i)-f(x_i)|\leqslant E_n(f)$, $k=1, 2$, la dernière relation n'a lieu que si

$$Q_n^1(x_i)-f(x_i)=Q_n^2(x_i)-f(x_i).$$

Deux polynômes distincts $Q_n^1(x)$ et $Q_n^2(x)$ de degré n coïncident en $n+2$ points, ce qui est contraire à l'hypothèse.

PROBLÈME. La fonction $f(x) = \sin 100x$ est approchée sur $[0, \pi]$. Trouver $Q_{90}(x)$.

§ 6. Exemples de meilleure approximation uniforme

1. Une fonction $f(x)$ continue sur $[a, b]$ est approchée par un polynôme de degré zéro. Soit

$$\sup_{[a, b]} f(x)=f(x_1)=M, \qquad \inf_{[a, b]} f(x)=f(x_2)=m.$$

$Q_0(x) = (M+m)/2$ est la meilleure approximation et x_1, x_2 les points du support de Tchebycheff.

PROBLÈMES. a) Démontrer que le polynôme de meilleure approximation de degré zéro est de la même forme si $f(x)$ n'est pas nécessairement continue.

b) Donner un exemple de fonction et de polynôme correspondant du premier degré, meilleure approximation uniforme, de sorte que les points a et b ne vérifient pas les conditions du théorème d'alternance de Tchebycheff.

c) Donner un exemple de fonction (non continue bien sûr) telle que son polynôme de meilleure approximation ne satisfasse pas aux conditions du théorème de Tchebycheff.

2. Une fonction $f(x)$ convexe sur $[a, b]$ est approchée par le polynôme du premier degré $Q_1(x) = a_0 + a_1x$. Du moment que $f(x)$ est convexe, la différence $f(x) - (a_0 + a_1x)$ n'a qu'un seul point d'extrémum intérieur. Aussi les points a et b appartiennent-ils au support de Tchebycheff. Soit d un troisième point de ce type. Selon le théorème de Tchebycheff, on a

$$\begin{aligned} f(a) - (a_0 + a_1a) &= \alpha L, \\ f(d) - (a_0 + a_1d) &= -\alpha L, \\ f(b) - (a_0 + a_1b) &= \alpha L. \end{aligned}$$

En retranchant la première équation de la troisième, il vient $f(b) - f(a) = a_1(b - a)$. D'où $a_1 = \frac{f(b)-f(a)}{b-a}$. Pour déterminer les inconnues d, L, a_0, a_1 et $\alpha = \pm 1$ on ne dispose que de trois équations. Rappelons toutefois que d est le point d'extrémum de la différence $f(x) - (a_0 + a_1 x)$. Trouvons d; si $f(x)$ est dérivable, on le fait à l'aide de l'équation $f'(d) - a_1 = 0$. Déterminons maintenant a_0, par exemple à partir de l'équation qu'on obtient en additionnant les deux premières. Géométriquement, on procède comme

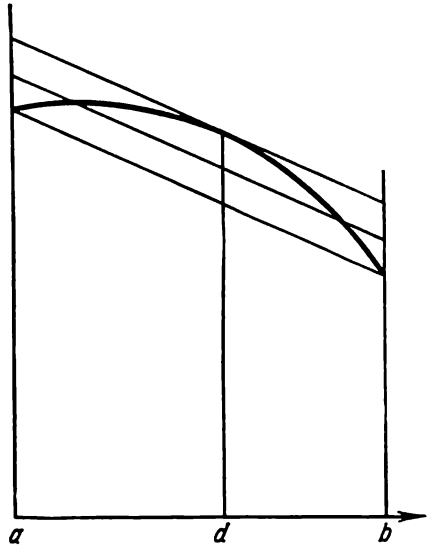


Fig. 4.6.1

suit (fig. 4.6.1.). Menons une sécante par les points $(a, f(a))$, $(b, f(b))$. Soit a_1 sa pente. Traçons une tangente à la courbe $y = f(x)$ de pente a_1 et une droite parallèle aux deux précédentes et située à égale distance d'elles.

PROBLÈME. $f(x) = |x|$, $[a, b] = [-1, 5]$. Construire le polynôme de meilleure approximation uniforme de degré 1.

3. La fonction $f(x)$ satisfaisant à la condition $f^{(n+1)}(x) \geqslant 0$ est approchée sur $[a, b]$ par un polynôme de meilleure approximation de degré n; on demande d'évaluer $E_n(f)$. Au § 9 du chapitre II, nous avons évalué l'erreur dans l'interpolation avec les points zéros du polynôme de Tchebycheff

$$x_k = \frac{a+b}{2} + \frac{b-a}{2} \cos\left(\frac{\pi(2k-1)}{2(n+1)}\right),$$

à savoir

$$|f(x)-P_n(x)|\leqslant\left(\max_{[a,\,b]}|f^{(n+1)}(x)|\right)\frac{(b-a)^{n+1}}{2^{2n+1}(n+1)!}. \qquad (1)$$

D'où l'inégalité

$$E_n(f)\leqslant\left(\max_{[a,\,b]}|f^{(n+1)}(x)|\right)\frac{(b-a)^{n+1}}{2^{2n+1}(n+1)!}.$$

Soit $Q_n(x)$ le polynôme de meilleure approximation uniforme. Selon le théorème de Tchebycheff la différence $f(x)-Q_n(x)$ s'annule au $(n+1)$-ième point $y_1, \ldots, y_{n+1}$. On peut donc dire que $Q_n(x)$ est un polynôme d'interpolation basé sur les points $y_1, \ldots, y_{n+1}$. D'après (2.3.1) on a pour l'erreur d'interpolation

$$f(x)-Q_n(x)=f^{(n+1)}(\zeta)\frac{\omega_{n+1}(x)}{(n+1)!},$$

où

$$\omega_{n+1}(x)=(x-y_1)\ldots(x-y_{n+1}),\ \zeta\in[a,\,b].$$

Supposons que

$$\max_{[a,\,b]}|\omega_{n+1}(x)|=|\omega_{n+1}(x_0)|.$$

On a

$$E_n(f)=\|f(x)-Q_n(x)\|\geqslant$$

$$\geqslant|f(x_0)-Q_n(x_0)|=|f^{(n+1)}(\zeta(x_0))|\frac{|\omega_{n+1}(x_0)|}{(n+1)!}\geqslant$$

$$\geqslant\left(\min_{[a,\,b]}|f^{(n+1)}(x)|\right)\left(\max_{[a,\,b]}\frac{|\omega_{n+1}(x)|}{(n+1)!}\right).$$

Conformément à (2.8.6)

$$\max_{[a,\,b]}|\omega_{n+1}(x)|\geqslant\frac{(b-a)^{n+1}}{2^{2n+1}},$$

d'où l'estimation

$$E_n(f)\geqslant\left(\min_{[a,\,b]}|f^{(n+1)}(x)|\right)\frac{(b-a)^{n+1}}{2^{2n+1}(n+1)!}. \qquad (2)$$

Ainsi, si $f^{(n+1)}(x)$ conserve le signe et ne varie pas trop, la différence entre l'erreur sur le polynôme de meilleure approximation et celle sur le polynôme d'interpolation pour les zéros des polynômes de Tchebycheff n'est pas essentielle.

PROBLÈME. Démontrer que dans le cas examiné les points a et b satisfont aux conditions du théorème de Tchebycheff.

4. Proposons-nous de trouver le polynôme de meilleure approximation de degré n dans le cas

$$f(x)=P_{n+1}(x)=a_0+\ldots+a_{n+1}x^{n+1},\ a_{n+1}\neq 0.$$

Alors $f^{(n+1)}(x) = a_{n+1}(n+1)!$, et la majoration (1) et la minoration (2) coïncident pour $E_n(f)$:

$$E_n(f) = |a_{n+1}|(b-a)^{n+1}2^{-2n-1}.$$

Ainsi, le meilleur approximant est le polynôme d'interpolation $Q_n(x)$ basé sur les abscisses

$$x_k = \frac{a+b}{2} + \frac{b-a}{2}\cos\left(\frac{\pi(2k-1)}{2(n+1)}\right), \quad k = 1, \ldots, n+1.$$

Ce polynôme de meilleure approximation uniforme peut être représenté d'une façon différente:

$$Q_n(x) = P_{n+1}(x) - a_{n+1}T_{n+1}\left(\frac{2x-(a+b)}{b-a}\right)\frac{(b-a)^{n+1}}{2^{2n+1}}. \qquad (3)$$

En effet, nous avons à droite un polynôme de degré n puisque le coefficient de x^{n+1} est nul; les points de maximum $x_i = \frac{a+b}{2} + \frac{b-a}{2}\cos\frac{\pi i}{n+1}$, $i = 0, \ldots, n$, du module de $P_{n+1} - Q_n(x)$ forment le support de Tchebycheff sur $[a, b]$.

5. Supposons que $f(x)$ est continue et impaire par rapport au milieu du segment $[a, b]$. Arrêtons-nous, pour fixer les idées, sur le segment $[-1, 1]$. La condition d'imparité signifie que $f(x) = -f(-x)$. Montrons que le polynôme de meilleure approximation de degré quelconque est ici impair, i.e. s'écrit comme somme de puissances impaires de x. En effet, soit $Q_n(x)$ un polynôme de meilleure approximation de $f(x)$. On a $|f(x) - Q_n(x)| \leqslant E_n(f)$. En remplaçant x par $-x$ et en multipliant le module par -1, il vient

$$|-f(-x) - f(-Q_n(-x))| \leqslant E_n(f),$$

ou

$$|f(x) - (-Q_n(-x))| \leqslant E_n(f).$$

Par conséquent, $-Q_n(-x)$ est lui aussi un polynôme de meilleure approximation uniforme. Vu le théorème d'unicité, $Q_n(x) = -Q_n(-x)$, c.q.f.d.

6. On demande d'approcher x^3 sur $[-1, 1]$ par le polynôme de meilleure approximation de degré 1. Le résultat précédent admet deux utilisations. 1) Le polynôme cherché étant impair, on le recherchera parmi les polynômes $Q_1(x) = a_1x$. 2) Puisque pour le problème proposé le polynôme de meilleure approximation du deuxième degré se trouve être du premier degré, le problème initial équivaut à construire un meilleur approximant de degré 2. Or, nous avons déjà considéré le problème de meilleure approximation d'un polynôme par un autre polynôme de degré inférieur d'une unité.

PROBLÈME. $f(x)$ est continue et paire par rapport au milieu du segment $[-1, 1]$ sur lequel on approche: $f(x) = f(-x)$.

Démontrer que le polynôme de meilleure approximation $Q_n(x)$ est pair.

PROBLÈME. Approcher $f(x) = \exp(x^2)$ sur $[-1, 1]$ par le polynôme de meilleure approximation du troisième degré.

Il résulte de ce qui précède que ce dernier est de la forme $a_0 + a_2x^2$. Le problème équivaut donc à approcher au mieux $f_1(y) = \exp y$ sur $[0, 1]$ par le polynôme $a_0 + a_2y$ (voir n° 2).

7. Un cas assez répandu est celui où l'on se contente d'une bonne approximation sans chercher la meilleure approximation uniforme d'une fonction. Par souci de la simplicité, bornons-nous à une approximation sur le segment $[-1, 1]$. Développons la fonction $f(x)$ en série suivant un système orthogonal de polynômes de Tchebycheff

$$f(x) \sim \sum_{j=0}^{\infty} d_j T_j(x).$$

Il arrive souvent qu'un morceau de cette série

$$\sum_{j=0}^{n} d_j T_j(x)$$

de degré peu élevé assure une bonne approximation uniforme.

Les coefficients d_j sont parfois difficiles à obtenir explicitement, mais on connaît en revanche le développement taylorien

$$f(x) = \sum_{j=0}^{\infty} a_j x^j,$$

qui converge pour $|x| \leqslant 1$. Dans ces conditions il serait bon de recourir à la méthode suivante dite *méthode télescopique*. On choisit un n tel que l'erreur résultant de la formule

$$f(x) \approx P_n(x) = \sum_{j=0}^{n} a_j x^j$$

est suffisamment petite. On approche ensuite $P_n(x)$ par le polynôme $P_{n-1}(x)$, sa meilleure approximation uniforme. Selon la formule (3)

$$P_{n-1}(x) = P_n(x) - a_n T_n(x)\, 2^{1-n}.$$

Puisque $|T_n(x)| \leqslant 1$ sur $[-1, 1]$, on a

$$|P_{n-1}(x) - P_n(x)| \leqslant |a_n|\, 2^{1-n},$$

puis on approche $P_{n-1}(x)$ par $P_{n-2}(x)$, sa meilleure approximation uniforme, et ainsi de suite, tant que l'erreur provenant de ces approximations successives reste faible.

Ce procédé admet une description différente. Développons $P_n(x)$ suivant les polynômes de Tchebycheff

$$P_n(x) = \sum_{j=0}^{n} d_j T_j(x).$$

Introduisons les notations $Q_m(x) = \sum_{j=0}^{m} d_j T_j(x)$ pour $m \leqslant n$. Tout polynôme $Q_m(x)$ est un polynôme de meilleure approximation uniforme de degré m pour $Q_{m+1}(x)$; ceci étant,

$$E_m(Q_{m+1}) = \| Q_{m+1} - Q_m \| = | d_{m+1} |. \tag{4}$$

Cela découle par exemple de la formule (3) ou directement du théorème de Tchebycheff. Il en résulte que $Q_{n-1}(x) = P_{n-1}(x)$, $Q_{n-2}(x) = P_{n-2}(x)$, etc.

L'idée de la méthode est la suivante. La fonction de départ est approchée par le terme $P_n(x)$ de sa série de Taylor, puis $P_n(x)$ est développé suivant les polynômes de Tchebycheff et on rejette plusieurs derniers termes du développement. Du moment que

$$| P_n(x) - P_m(x) | \leqslant \sum_{j=m+1}^{n} | d_j |,$$

l'estimation totale de l'erreur s'écrit

$$| f(x) - P_m(x) | \leqslant \max | f(x) - P_n(x) | + \sum_{j=m+1}^{n} | d_j |.$$

Soit à approcher arctg x sur $\left[-\operatorname{tg}\frac{\pi}{8}, \operatorname{tg}\frac{\pi}{8} \right]$ à $0{,}5 \cdot 10^{-5}$ près. Une telle précision lors de l'approximation par un morceau de la série de Taylor implique

$$\operatorname{arctg} x \approx x - \frac{x^3}{3} + \frac{x^5}{5} - \frac{x^7}{7} + \frac{x^9}{9} - \frac{x^{11}}{11}.$$

En approchant le polynôme par le polynôme de meilleure approximation uniforme de degré inférieur d'une unité, on obtient au bout de trois pas:

$$P_5(x) = 0{,}9999374x - 0{,}3303433x^3 + 0{,}1632823^5.$$

Ce polynôme assure lui aussi la précision demandée. Notons que l'imparité du polynôme initial abaisse à chaque pas l'exposant de 2.

8. Pour évaluer $E_n(f)$ on préfère parfois au théorème de Vallée-Poussin le résultat suivant.

THÉORÈME. *Si* $f(x)$ *est approchée sur* $[-1, 1]$ *par*

$$Q_{n+1}(x) = a_{n+1}x^{n+1} + \ldots + a_0,$$

alors

$$E_n(f) \geqslant | a_{n+1} | 2^{-n} - \| f - Q_{n+1} \|.$$

DÉMONSTRATION. Si Q_m^f est la meilleure approximation de la fonction f et Q_m^g celle de la fonction g, alors

$$| (Q_m^f(x) + Q_m^g(x)) - (f(x) + g(x)) | \leqslant$$
$$\leqslant | Q_m^f(x) - f(x) | + | Q_m^g(x) - g(x) | \leqslant E_m(f) + E_m(g).$$

Il en découle

$$E_m(f+g) \leqslant E_m(f) + E_m(g).$$

Posant $m = n$, $g = Q_{n+1} - f$, il vient

$$E_n(Q_{n+1}) \leqslant E_n(f) + E_n(Q_{n+1} - f).$$

Compte tenu de (3), on a

$$E_n(Q_{n+1}) = |a_{n+1}| 2^{-n}.$$

Donc,

$$E_n(f) \geqslant E_n(Q_{n+1}) - E_n(Q_{n+1} - f) \geqslant$$
$$\geqslant |a_{n+1}| 2^{-n} - \|f - Q_{n+1}\|. \qquad (5)$$

Le théorème est démontré.

Voici l'exemple de $E_n(f)$ encadrée. Soit

$$f(x) = \sum_{j=0}^{\infty} d_j T_j(x);$$

posons

$$Q_m(x) = \sum_{j=0}^{m} d_j T_j(x).$$

On a de toute évidence

$$\|f - Q_{n+1}\| \leqslant \sum_{j=n+2}^{\infty} |d_j|.$$

Conformément à (4), $E_n(Q_{n+1}) = |d_{n+1}|$. Avec la première inégalité (5) on a

$$E_n(f) \geqslant |d_{n+1}| - \sum_{j=n+2}^{\infty} |d_j|.$$

D'autre part, on a évidemment

$$E_n(f) \leqslant \|f - Q_n\| \leqslant \sum_{j=n+1}^{\infty} |d_j| = |d_{n+1}| + \sum_{j=n+2}^{\infty} |d_j|.$$

Pour $|d_{n+1}| > \sum_{j=n+2}^{\infty} |d_j|$, la comparaison de ces estimations donne l'encadrement de $E_n(f)$.

§ 7. Méthode itérative de construction du polynôme de meilleure approximation uniforme

On recherche d'ordinaire le polynôme de meilleure approximation uniforme par une méthode itérative quelconque. L'algorithme décrit ci-dessous consiste à trouver par approximations successives

les points du support de Tchebycheff et les polynômes associés qui convergent vers le polynôme de meilleure approximation uniforme.

On suppose dans l'exposé qui suit que la fonction à approcher est continue et n'est pas un polynôme de degré n, i.e. $E_n(f) \neq 0$.

Supposons que les points vérifiant les conditions du théorème de Tchebycheff sont représentés approximativement par $x_0^k < \ldots < x_{n+1}^k$. Construisons le polynôme

$$Q_n^k(x) = \sum_{j=0}^{n} a_j^k x^j,$$

qui satisfait en ces points aux relations

$$f(x_i^k) - Q_n^k(x_i^k) = (-1)^i E^k \tag{1}$$

avec E^k inconnu. On aboutit aux équations

$$\sum_{j=0}^{n} a_j^k (x_i^k)^j + (-1)^i E^k = f(x_i^k) \tag{2}$$

par rapport aux $n+2$ inconnues $a_0^k, \ldots, a_n^k, E^k$.

Montrons que le déterminant de ce système est non nul. Raisonnons par l'absurde; le système homogène

$$Q_n^k(x_i) + (-1)^i E^k = \sum_{j=0}^{n} a_j^k (x_i^k)^j + (-1)^i E^k = 0$$

a alors une solution non nulle. Si pour cette solution $E^k = 0$, l'un au moins des a_j^k est différent de zéro : le polynôme non nul $Q_n^k(x)$ de degré n possède $n+2$ zéros, et l'on aboutit à une contradiction. Si $E^k \neq 0$, l'un au moins des a_j^k est également non nul et le polynôme non nul $Q_n^k(x)$ change de signe entre n'importe quels deux points x_i^k, x_{i+1}^k, c'est-à-dire qu'il a au moins $n+1$ zéros, ce qui nous conduit une fois de plus à une contradiction.

Notons

$$x_{-1}^k = a, \quad x_{n+2}^k = b, \quad R_n^k(x) = f(x) - Q_n^k(x).$$

Au cours des itérations on fixe un certain q_0 dans les limites $0 < q_0 < 1$.

Si $k > 0$, on identifie aux x_i^k les points appartenant aux segments $[x_{i-1}^{k-1}, x_{i+1}^{k-1}]$, $i = 0, \ldots, n+1$, et vérifiant dans l'ensemble les conditions

$$\alpha) \quad \max_{0 \leqslant i \leqslant n+1} |R_n^{k-1}(x_i^k)| \geqslant |E^{k-1}| + q_0 (\|R_n^{k-1}(x)\| - |E^{k-1}|),$$

$$\beta) \quad R_n^{k-1}(x_i^k) R_n^{k-1}(x_{i+1}^k) \leqslant 0 \qquad \text{pour} \quad i = 0, \ldots, n,$$

$$\gamma) \quad |R_n^{k-1}(x_i^k)| \geqslant |E^{k-1}| \quad \text{pour} \quad i = 0, \ldots, n+1.$$

L'ensemble de ces points n'est pas vide et peut être construit comme suit.

Prenons un point x^k quelconque tel que

$$|R_n^{k-1}(x^k)| \geqslant |E^{k-1}| + q_0 (\|R_n^{k-1}(x)\| - |E^{k-1}|);$$

vu que le second membre de cette expression ne dépasse pas $\| R_n^{k-1}(x) \|$, ce point existe bien. Si $E^{k-1} = 0$, les conditions imposées sont vérifiées par un ensemble de x^k et de $n+1$ points quelconques parmi $x_0^{k-1}, \ldots, x_{n+1}^{k-1}$.

Pour $E^{k-1} \neq 0$, tous les $| R_n^{k-1}(x_i^{k-1}) | \neq 0$; considérons les cas suivants:

a) $$x^k \in [x_{-1}^{k-1}, x_0^{k-1}],$$

b) $$x^k \in [x_{n+1}^{k-1}, x_{n+2}^{k-1}],$$

c) $$x^k \in [x_0^{k-1}, x_{n+1}^{k-1}].$$

Dans a) posons $x_0^k = x^k$,

$$x_i^k = \begin{cases} x_i^{k-1} & \text{pour } R_n^{k-1}(x_0^{k-1}) R_n^{k-1}(x_0^k) > 0, \\ x_{i-1}^{k-1} & \text{pour } R_n^{k-1}(x_0^{k-1}) R_n^{k-1}(x_0^k) < 0, \end{cases} \quad i = 1, \ldots, n+1.$$

Dans b) posons $x_{n+1}^k = x^k$,

$$x_i^k = \begin{cases} x_i^{k-1} & \text{pour } R_n^{k-1}(x_{n+1}^{k-1}) R_n^{k-1}(x_{n+1}^k) > 0, \\ x_{i+1}^{k-1} & \text{pour } R_n^{k-1}(x_{n+1}^{k-1}) R_n^{k-1}(x_{n+1}^k) < 0, \end{cases} \quad i = 0, \ldots, n.$$

Plaçons-nous maintenant dans le cas c); soit

$$x^k \in [x_{i_0}^{k-1}, x_{i_0+1}^{k-1}], \quad 0 \leqslant i_0 \leqslant n;$$

si $R_n^{k-1}(x^k) R_n^{k-1}(x_{i_0}^{k-1}) > 0$, posons

$$x_{i_0}^k = x^k, \quad x_i^k = x_i^{k-1}, \quad i \neq i_0;$$

si $R_n^{k-1}(x^k) R_n^{k-1}(x_{i_0+1}^{k-1}) > 0$, posons

$$x_{i_0+1}^k = x^k, \quad x_i^k = x_i^{k-1}, \quad i \neq i_0+1.$$

Dans tous les cas les ensembles de points x_i^k obtenus vérifient les conditions $\alpha)$, $\beta)$, $\gamma)$.

En l'absence de toute information complémentaire sur la position des points du support de Tchebycheff qu'on cherche, on prend souvent pour x_i^0 les points d'extrémum du polynôme de Tchebycheff

$$x_i^0 = \frac{a+b}{2} + \frac{b-a}{2} \cos\left(\pi \frac{n+1-i}{n+1}\right), \quad i = 0, \ldots, n+1.$$

PROBLÈME. Soit $f(x)$ un polynôme de degré $n+1$. Montrer que pour de pareilles x_i^0, $Q_n^0(x)$ s'avère le polynôme de meilleure approximation uniforme cherché.

Si la fonction $f(x)$ est près d'être paire ou impaire par rapport au milieu du segment, il serait préférable de prendre pour x_i^0 les points d'extrémum des polynômes de Tchebycheff de degré correspondant. S'il faut construire les polynômes de meilleure approximation pour une grande série de fonctions $f(A, x)$ dépendant du paramètre A, on recommande parfois le procédé suivant

sous réserve que les points x_i vérifiant les hypothèses du théorème de Tchebycheff soient des fonctions régulières en A : après avoir trouvé ces points pour certains $A_1, \ldots, A_m$, on détermine par interpolation les premières approximations pour d'autres A.

THÉORÈME. *Soit*

$$\max_{[a,\,b]} |f^{(l)}(x)| = M_l < \infty, \quad l = 0, \ldots, n+2.$$

Alors pour le processus itératif décrit, on a

$$\|f - Q_n^k\| \leqslant E_n(f) + \text{const}\, q^k$$

pour $q < 1$.

Ceci étant, $\|Q_n^k - Q_n\| \to 0$, *où* Q_n *est le polynôme de meilleure approximation uniforme.*

DÉMONSTRATION. La différence $f - Q_n^k$ s'annule en certains points $y_1^k, \ldots, y_{n+1}^k$ par suite du changement de signe lorsqu'on passe du point x_i^k au point x_{i+1}^k. C'est pourquoi $Q_n^k(x)$ est un polynôme d'interpolation pour $f(x)$ et est basé sur $y_1^k, \ldots, y_{n+1}^k \in [a, b]$. Selon (2.3.1) et (2.17.5), on a pour $x \in [a, b]$

$$R_n^k(x) = f(x) - Q_n^k(x) = f^{(n+1)}(\zeta_1)\frac{\Omega^k(x)}{(n+1)!},$$

$$(R_n^k(x))' = f'(x) - (Q_n^k(x))' = f^{(n+1)}(\zeta_1)\frac{(\Omega^k(x))'}{(n+1)!} + f^{(n+2)}(\zeta_2)\frac{\Omega^k(x)}{(n+2)!};$$

ici $\Omega^k(x) = (x - y_1^k) \ldots (x - y_{n+1}^k)$, $\zeta_1, \zeta_2 \in [a, b]$. Puisque $|(\Omega^k(x))'| \leqslant (n+1) \times \times (b-a)^n$, on aboutit à l'estimation

$$|R_n^k(x)| \leqslant M_{n+1}\frac{(b-a)^{n+1}}{(n+1)!} = D_1,$$

$$|(R_n^k(x))'| \leqslant M_{n+1}\frac{(b-a)^n}{n!} + M_{n+2}\frac{(b-a)^{n+1}}{(n+2)!} = D_2$$

quels que soient k.

Mettons le système de relations (1) sous la forme

$$R_n^k(x_i^k) = E^k \psi^k(x_i^k),$$

où $\psi^k(x_i^k) = (-1)^i$ en x_i^k. Formons la différence divisée basée sur les abscisses $x_0^k, \ldots, x_{n+1}^k$ des deux membres :

$$R_n^k(x_0^k; \ldots; x_{n+1}^k) = E^k \psi^k(x_0^k; \ldots; x_{n+1}^k).$$

Vu que la différence divisée d'un polynôme de degré n aux $(n+2)$ abscisses est nulle, on remplace R_n^k par R_n^{k-1} ; il vient

$$R_n^{k-1}(x_0^k; \ldots; x_{n+1}^k) = E^k \psi^k(x_0^k; \ldots; x_{n+1}^k),$$

ou

$$E^k = \frac{R_n^{k-1}(x_0^k; \ldots; x_{n+1}^k)}{\psi^k(x_0^k; \ldots; x_{n+1}^k)}.$$

En ouvrant les parenthèses on trouve

$$E^k = \sum_{i=0}^{n+1} \frac{d_i^k R_n^{k-1}(x_i^k)}{\sum_{i=0}^{n+1} (-1)^i d_i^k} \tag{3}$$

avec

$$d_i^k = \prod_{j \neq i} (x_i^k - x_j^k)^{-1}.$$

Les quantités $R_n^{k-1}(x_i^k)$, $i=0, \ldots, n+1$, k fixe, alternent en signe par suite du choix des points x_i^k ; les signes de d_i^k alternent eux aussi. Et la relation (3) s'écrit

$$|E^k| = \sum_{i=0}^{n+1} h_i^k |R_n^{k-1}(x_i^k)| \tag{4}$$

avec

$$h_i^k = |d_i^k| / \Big(\sum_{i=0}^{n+1} |d_i^k|\Big) > 0.$$

Il est évident que $\sum_{i=0}^{n+1} h_i^k = 1$. Si les x_i^k vérifient la condition γ), il vient

$$|R_n^{k-1}(x_i^k)| \geqslant |E^{k-1}|, \quad i=0, \ldots, n+1.$$

Donc (4) fournit

$$|E^k| \geqslant \sum_{i=0}^{n+1} h_i^k |E^{k-1}| = |E^{k-1}|. \tag{5}$$

Si Q_n^{k-1} n'est pas la meilleure approximation uniforme cherchée, les points $x_0^{k-1}, \ldots, x_{n+1}^{k-1}$ ne vérifient pas les conditions du théorème de Tchebycheff et l'on a, conformément à α),

$$|R_n^{k-1}(x_i^k)| > |E^{k-1}|,$$

pour au moins un i dans l'intervalle $0 \leqslant i \leqslant n+1$. Alors (5) fait place à $|E^k| > |E^{k-1}|$. Il se peut que $E^0 = 0$, mais compte tenu de la dernière inégalité, on a, pour tout $k \geqslant 1$, $|E^k| \geqslant |E^1| > |E^0| \geqslant 0$. Aussi, quels que soient $k \geqslant 1$ et $0 \leqslant i \leqslant n$, on peut écrire la suite de relations

$$2|E^1| \leqslant 2|E^k| = |R_n^k(x_{i+1}^k) - R_n^k(x_i^k)| \leqslant$$

$$\leqslant \max_{[a,\, b]} |(R_n^k(x))'| (x_{i+1}^k - x_i^k) \leqslant D_2 (x_{i+1}^k - x_i^k).$$

Pour $k \geqslant 1$, $0 \leqslant i < j \leqslant n+1$, il s'ensuit

$$x_j^k - x_i^k \geqslant x_{i+1}^k - x_i^k \geqslant \varepsilon > 0, \quad \varepsilon = 2\frac{|E^1|}{D_2} ;$$

vu que, de plus, $|x_i^k - x_j^k| \leqslant b - a$, il vient

$$(b-a)^{-(n+1)} \leqslant |d_i^k| \leqslant \varepsilon^{-(n+1)}.$$

Il en résulte

$$h_i^k \geqslant \frac{(b-a)^{-(n+1)}}{\sum_{i=0}^{n+1} \varepsilon^{-(n+1)}} = \frac{1}{n+2}\left(\frac{\varepsilon}{b-a}\right)^{n+1} = d, \tag{6}$$

d ne dépendant pas de k et i. Si α) est satisfaite, il existe au moins un p, $0 \leqslant p \leqslant n+1$, tel que

$$|R_n^{k-1}(x_p^k)| \geqslant |E^{k-1}| + q_0(\|R_n^{k-1}\| - |E^{k-1}|).$$

La relation (4), compte tenu de (6), donne

$$|E^k| \geqslant \sum_{i \neq p} h_i^k |E^{k-1}| + h_p^k(|E^{k-1}| + q_0(\|R_n^{k-1}\| - |E^{k-1}|)) =$$

$$= |E^{k-1}| + h_p^k q_0(\|R_n^{k-1}\| - |E^{k-1}|) \geqslant |E^{k-1}| + dq_0(\|R_n^{k-1}\| - |E^{k-1}|). \tag{7}$$

Puisque $E_n(f) \leqslant \|R_n^{k-1}\|$,

$$E_n(f) - |E^k| \leqslant (1 - dq_0)(E_n(f) - |E^{k-1}|),$$

et

$$0 \leqslant E_n(f) - |E^k| \leqslant (1-dq_0)^k (E_n(f) - |E^0|), \quad |E^k| = E_n(f) + O(q^k).$$

(7) entraîne également

$$\|R_n^{k-1}\| - |E^k| \leqslant (dq_0)^{-1}(|E^k| - |E^{k-1}|);$$

d'où

$$\|R_n^k\| = E_n(f) + O(q^k). \tag{8}$$

Ce qui démontre la première proposition du théorème.

Lorsqu'on veut, dans la pratique, rechercher le polynôme de meilleure approximation, on effectue d'habitude un déplacement complémentaire à partir des points x_i^k proposés plus haut afin que chacune des quantités $|R^{k-1}(x_i^k)|$ soit la plus grande possible et que les conditions α), β), γ) soient réalisées.

Passons à la deuxième proposition du théorème. Notons que la convergence de $Q_n^k(x)$ vers $Q_n(x)$ ne possède en fait aucun intérêt pratique. En effet, si l'on construit le polynôme de meilleure approximation, c'est pour calculer à son aide les valeurs de la fonction. Si, au lieu de ce polynôme, l'on obtient $Q_n^k(x)$ pour lequel $\|Q_n^k - f\| \approx E_n(f) = \|Q_n - f\|$, le degré de proximité de $Q_n^k(x)$ et $Q_n(x)$ est sans rapport avec le but à atteindre.

Fixons $n+1$ points $x_0, \ldots, x_n \in [a, b]$ distincts. Soit S^k un point de coordonnées $(Q_n^k(x_0), \ldots, Q_n^k(x_n))$. Montrons que

$$S^k \to S = (Q_n(x_0), \ldots, Q_n(x_n)) \text{ pour } k \to \infty.$$

Supposons le contraire, i.e. qu'il existe une suite partielle $k(j) \to \infty$ pour laquelle

$$\rho(S^{k(j)}, S) \geqslant \gamma > 0 \tag{9}$$

quel que soit j. ρ est la distance euclidienne usuelle. Les quantités $|Q_n^k(x_i)|$ sont uniformément bornées par suite de l'inégalité

$$|f(x) - Q_n^k(x)| \leqslant D_1.$$

Par conséquent, l'ensemble des points $S^{k(j)}$ est contenu dans une portion finie de l'espace à $(n+1)$ dimensions et est un compact. On peut donc isoler de

la suite $S^{k(j)}$ une suite partielle $S^{k(j(m))}$ convergente vers une limite

$$S_0 = (b_0, \ldots, b_n) \text{ lorsque } m \to \infty.$$

Le polynôme d'interpolation de degré n pour $Q_n^k(x)$ et les points $x_0, \ldots, x_n$ coïncide avec lui-même; en recourant à la formule du polynôme d'interpolation de Lagrange, on a

$$Q_n^k(x) = \sum_{i=0}^{n} Q_n^k(x_i)\,\omega_i(x) \tag{10}$$

avec

$$\omega_i(x) = \prod_{j \neq i} \frac{x - x_j}{x_i - x_j}.$$

Puisque $Q_n^{k(j(m))} \to b_i$ pour $m \to \infty$, la relation (10) donne

$$Q_n^{k(j(m))} \to \sum_{i=0}^{n} b_i \omega_i(x) = \widetilde{Q}_n(x) \tag{11}$$

uniformément sur $[a, b]$ pour $m \to \infty$. On a

$$\| \widetilde{Q}_n - f \| \leqslant \| \widetilde{Q}_n - Q_n^{k(j(m))} \| + \| Q_n^{k(j(m))} - f \|.$$

Lorsque $m \to \infty$, le premier terme tend vers zéro d'après (11), et le deuxième vers $E_n(f)$, comme le veut (8). Aussi

$$\| \widetilde{Q}_n - f \| \leqslant E_n(f).$$

Le théorème d'unicité du polynôme de meilleure approximation uniforme entraîne

$$\widetilde{Q}_n(x) = Q_n(x).$$

Par conséquent,

$$Q_n^{k(j(m))}(x_i) \to \widetilde{Q}_n(x_i) = Q_n(x_i)$$

quel que soit i. Cela signifie que $S^{k(j(m))} \to S$, ce qui est contraire à (9). Ainsi, on vient d'établir que $S^k \to S$ pour $k \to \infty$. (11) entraîne alors que $Q_n^k(x) \to Q_n(x)$ uniformément sur $[a, b]$. C.Q.F.D.

Pourquoi prendre $q_0 < 1$? pourra-t-on demander. Est-ce que $q_0 = 1$ n'est pas la meilleure valeur? L'estimation (8) serait alors plus précise et, à ce qu'il paraît, le processus itératif convergerait plus rapidement. La question du choix de q_0 pour chaque classe concrète doit être tranchée par des calculs expérimentaux. Dans cet algorithme le plus gros travail est nécessité par la résolution du système (2) et le choix des points x_i^k. Ce choix dépend à son tour de la complexité du calcul des valeurs de $f(x)$. Si ces dernières se calculent sans trop de difficulté, il est tout indiqué de prendre le plus grand q_0. Il ne faut pas non plus perdre de vue qu'aux premières étapes de l'approximation nous sommes loin de la solution cherchée et certaine détérioration de la précision de l'algorithme est donc permise qui allège les premières itérations.

Ainsi, on modifie parfois l'algorithme décrit plus haut. A cette fin on fait correspondre à tout m naturel un w_m: $0 < w_m < 1$, et l'ensemble M_m des points du segment $[a, b]$: $y_1^m, \ldots, y_{N(m)}^m$. On exige de plus que pour $m \to \infty$ le nombre de points y_j^m tombant dans tout segment $[a_1, b_1] \subset [a, b]$ tende vers

l'infini et que $w_m \to 1$. L'algorithme qui permet de trouver d'une façon approchée le polynôme de meilleure approximation consiste à construire l'un après l'autre des polynômes $P_n^m(x)$ vérifiant les conditions suivantes: on choisit, parmi les points y_i^m, $n+2$ points $y_{i_0}^m < \ldots < y_{i_{n+1}}^m$ tels que les signes de $f(y_{i_k}^m) - P_n^m(y_{i_k}^m)$ diffèrent aux points suivants et que

$$\min_{0 \leqslant k \leqslant n+1} |f(y_{i_k}^m) - P_n^m(y_{i_k}^m)| \geqslant w_m \max_{1 \leqslant i \leqslant N(m)} |f(y_i^m) - P_n^m(y_i^m)|.$$

Le polynôme $P_n^m(x)$ est le point de départ des approximations du polynôme $P_n^{m+1}(x)$. L'algorithme itératif de recherche de $P_n^m(x)$ ressemble exactement à l'algorithme de calcul du polynôme de meilleure approximation uniforme pour les polynômes considérés sur tout le segment $[a, b]$. La seule différence est que les points x_i^k sont choisis parmi les éléments de l'ensemble M_m. Un cas assez répandu est celui où M_m est l'ensemble des points de maximum des polynômes de Tchebycheff M_m^0: $y_i^m = \frac{a+b}{2} + \frac{b-a}{2} \cos\left(\pi \frac{i-1}{N(m)-1}\right)$. Généralement $N(m) = sp^m + 1$ (s, p entiers); alors $M_m^0 \subset M_{m+1}^0$; les valeurs utilisées le plus souvent sont $s = n+1$ et $p = 2$.

On remarquera que notre description du programme standard de recherche du polynôme de meilleure approximation uniforme est loin d'être exhaustive. En particulier, il est possible de déterminer x_i^k soit à l'aide du tableau de toutes les valeurs de $f(x) - Q_n^k(x)$ aux points de l'ensemble M_m, soit par un procédé plus rapide mais moins sûr; nous n'avons pas non plus indiqué comment choisir les w_m.

§ 8. Ecriture du polynôme

Citons le cas d'un programme général de meilleure approximation uniforme qu'on a mis au point pour les approximations par des polynômes de degré peu élevé, puis introduit dans la bibliothèque de programmes généraux. Lorsqu'on a essayé d'opérer avec des polynômes de degré élevé, le programme s'est parfois arrêté sans donner l'approximation voulue ou bien s'est déroulé trop longtemps, et a entraîné la suspension des calculs.

On a d'abord pensé que l'échec était dû à une grande erreur de calcul commise en résolvant le système d'équations (7.2). On a donc abandonné le calcul direct et déterminé E^k à partir de (7.3). On a obtenu ainsi, compte tenu de (7.2), les valeurs de $Q_n^k(x_i^k)$, et les valeurs de $Q_n^k(x)$ nécessaires à l'approximation suivante ont été calculées d'après la formule d'interpolation de Lagrange

$$Q_n^k(x) = \sum_{i=0}^{n} Q_n^k(x_i^k)\, \omega_i(x), \quad \omega_i(x) = \prod_{j \neq i,\ 0 \leqslant j \leqslant n} \frac{x - x_j^k}{x_i^k - x_j^k}. \tag{1}$$

Le processus terminé, le dernier des polynômes obtenus, à savoir $P_n(x) = Q_n^k(x)$, a été écrit sous forme habituelle

$$P_n(x) = a_0 + a_1 x + \ldots + a_n x^n$$

puis on a fait imprimer les valeurs suivantes:

$$\{a_j\} \text{ et } e_n = \| P_n(x) - f(x) \|.$$

Cette solution s'est cependant avérée insatisfaisante.

On a constaté dans plusieurs cas une divergence sensible entre les valeurs de $f(x)$ et les $P_n(x)$ calculés par le schéma de Hörner:

$$P_n(x) = a_0 + x(a_1 + x(a_2 + \ldots + x(a_{n-1} + xa_n)\ldots)).$$

Cela tenait à ce que les $|a_j|$ étaient très grands et affectés donc d'une erreur absolue importante.

Le fait est que pour nombre de fonctions la somme $A_n = \sum_{j=0}^{n} |a_j|$ croît inévitablement si l'approximation de la fonction par des polynômes devient plus précise. On montrera plus bas que la vitesse de cette croissance est fonction des propriétés analytiques de la fonction à approximer.

Supposons qu'il y a une suite de polynômes

$$P^m(x) = \sum_{j=0}^{n(m)} a_j^m x^j,$$

tels que

$$|f(x) - P^m(x)| \leqslant 2^{-m} \text{ sur } [-1, 1], \tag{2}$$

$P^m(x)$ n'étant pas nécessairement les polynômes de meilleure approximation uniforme. Supposons encore que les coefficients de ces polynômes n'augmentent pas très vite:

$$l_m = \sum_{j=0}^{n(m)} |a_j^m| \leqslant M 2^{qm}, \quad 0 \leqslant q, \quad 1 \leqslant M. \tag{3}$$

THÉORÈME. *Si une suite de polynômes $P^m(x)$ vérifie les relations* (2), (3), *alors la fonction $f(x)$ est prolongeable analytiquement dans le domaine ouvert de la variable complexe G_{φ_0}, $\varphi_0 = \pi \frac{2q+1}{2q+2}$* (ce domaine est défini plus bas).

DÉMONSTRATION. Considérons les polynômes

$$Q^m(z) = P^m(z) - P^{m-1}(z);$$

par suite de (2) on a

$$|Q^m(x)| \leqslant |f(x) - P^m(x)| + |f(x) - P^{m-1}(x)| \leqslant 3 \cdot 2^{-m},$$

$$|P^m(z)| \leqslant \sum_{j=0}^{n(m)} |a_j^m| = l_m \quad \text{pour} \quad |z| \leqslant 1.$$

Donc

$$|Q^m(z)| \leqslant |P^m(z)| + |P^{m-1}(z)| \leqslant l_{m-1} + l_m \leqslant 2M 2^{qm}, \quad |z| \leqslant 1.$$

LEMME. *Si $g(z)$ est analytique dans le cercle $|z| \leqslant 1$ avec $|g(z)| \leqslant M_2$ sur le contour du cercle et $|g(z)| \leqslant M_1$ sur le segment $[-1, 1]$, alors dans ce cercle*

$$|g(z)| \leqslant M_1^{(2/\pi)\varphi(z) - 1} M_2^{2 - (2/\pi)\varphi(z)}, \tag{4}$$

$\varphi(z)$ *étant l'angle (non orienté) sous lequel on voit le segment* $[-1, 1]$ *du point* $z = x + iy$.

Démontrons le lemme pour les points du demi-cercle K_1: $\operatorname{Im} z \geqslant 0$. Considérons la fonction

$$w(z) = \left(\frac{2}{\pi}\varphi(z) - 1\right) \ln M_1 + \left(2 - \frac{2}{\pi}\varphi(z)\right) \ln M_2 - \ln |g(z)|.$$

$\varphi(z) = \pi$ sur le segment $[-1, 1]$, donc

$$w = \ln M_1 - \ln |g(z)| \geqslant 0; \tag{5}$$

$\varphi(z) = \frac{\pi}{2}$ sur l'arc de cercle, donc

$$w = \ln M_2 - \ln |g(z)| \geqslant 0. \tag{6}$$

Vu les égalités $\ln |g(z)| = \operatorname{Re} \ln g(z)$, on a

$$\varphi(z) = \arg\left(\frac{z-1}{z+1}\right) = \operatorname{Im} \ln\left(\frac{z-1}{z+1}\right),$$

les fonctions $\ln |g(z)|$ et $\varphi(z)$ étant harmoniques par rapport aux variables x, y; par conséquent, $w(x, y)$ est harmonique elle aussi, et sa borne inférieure

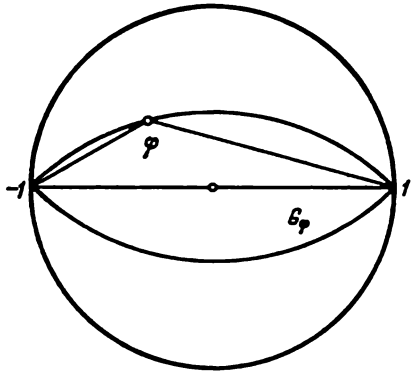


Fig. 4.8.1

dans K_1 est atteinte sur le contour. Compte tenu de (5) et (6), cela équivaut à $w(x, y) \geqslant 0$ dans K_1. D'où

$$\ln |g(z)| \leqslant \left(\frac{2}{\pi}\varphi(z) - 1\right) \ln M_1 + \left(2 - \frac{2}{\pi}\varphi(z)\right) \ln M_2.$$

En se débarrassant des logarithmes, on retrouve (4).

Appliqué à l'estimation $|Q^m(z)|$ le lemme donne

$$|Q^m(z)| \leqslant (3 \cdot 2^{-m})^{((2/\pi)\varphi - 1)} (2M \cdot 2^{qm})^{2-(2/\pi)\varphi} \leqslant 3M 2^{\left((2q+1)-(2q+2)\frac{\varphi}{\pi}\right)m}. \tag{7}$$

Désignons par G_φ un domaine ouvert limité par des arcs de cercle d'où l'on voit le segment $[-1, 1]$ sous l'angle φ (voir fig. 4.8.1). Puisque la fonction $2q + 1 - (2q + 2)\frac{\varphi}{\pi}$ décroît avec la croissance de φ et que de tous les points du domaine

G_φ le segment $[-1, 1]$ est vu sous un angle égal au moins à φ, l'inégalité (7) est vérifiée en tous les points de G_φ. Soit φ_0 une solution de l'équation

$$(2q+1)-(2q+2)\frac{\varphi_0}{\pi}=0, \tag{8}$$

c.-à-d. que $\varphi_0=\pi\dfrac{2q+1}{2q+2}$. Prenons un $\varphi_1>\varphi_0$ arbitraire. On a immédiatement

$$|P^m(z)|\leqslant|P^0(z)|+\sum_{k=1}^{m}|P^k(z)-P^{k-1}(z)|=|P^0(z)|+\sum_{k=1}^{m}|Q^k(z)|.$$

Conformément à (7), on a pour tous les points du domaine $\overline{G}_{\varphi_1}$:

$$|P^m(z)|\leqslant\max_{\overline{G}_{\varphi_1}}|P^0(z)|+\sum_{k=1}^{\infty}3M\cdot 2^{\left(2q+1-(2q+2)\frac{\varphi_1}{\pi}\right)m}=M(\varphi_1).$$

$M(\varphi_1)$ est fini, car pour $\varphi_1<\varphi_0$ le coefficient de m de l'exposant est négatif. Ainsi, pour $z\in\overline{G}_{\varphi_1}$ l'estimation $|P^m(z)|\leqslant|M(\varphi_1)|$ est uniforme en m. Rappelons une assertion bien connue. Soient $f_m(z)$ des fonctions analytiques dans le domaine ouvert G simplement connexe et uniformément bornées en module: $|f_m(z)|\leqslant C$. Supposons que la suite de valeurs de $f_m(z)$ converge sur un ensemble de valeurs de $z\in G$ possédant un point d'accumulation $z_0\in G$. La suite $f_m(z)$ converge alors vers une fonction $f(z)$ analytique dans G, $|f(z)|\leqslant C$. Cela signifie, dans notre cas, que la suite de polynômes $P^m(z)$ converge uniformément vers une certaine fonction $h(z)$ analytique dans G_{φ_1}. Il en résulte que la fonction $f(x)$ définie sur $[-1, 1]$ est prolongeable en tant que fonction analytique à tous les points du domaine ouvert G_{φ_1}. Etant donné que $\varphi_1>\varphi_0$ est quelconque, elle est également prolongeable à tous les points du domaine ouvert G_{φ_0}. Le théorème est démontré.

Quels sont les résultats pratiques de ce théorème? Si le nombre a est écrit en virgule flottante avec t chiffres binaires, son erreur peut dépasser $|a|2^{-t-1}$; au pis aller, si les coefficients a_j^m donnent lieu à des erreurs de même signe, l'erreur sur la valeur

$$P^m(1)=\sum_{j=0}^{n(m)}a_j^m$$

qui résulte des précédentes, peut être supérieure à

$$\left(\sum_{j=0}^{n(m)}|a_j|\right)2^{-t-1}=l_m2^{-t-1}.$$

Pour se faire une idée de l'influence qualitative de cette erreur, supposons que le nombre m figurant dans (2), (3) et t sont grands. Soit $f(x)$ une fonction analytique dans un domaine G_{φ_0} et qui ne l'est dans aucun G_φ pour $\varphi<\varphi_0$. Admettons que q_0 correspond à ce φ_0 selon (8), i.e. $q_0=\dfrac{2\varphi_0-\pi}{2(\pi-\varphi_0)}$. Prenons un $q<q_0$ quelconque. En supposant l'inégalité $l_m\leqslant 2^{qm}$ vérifiée pour tout m, en vertu

du théorème, $f(z)$ est analytique dans un G_φ correspondant pour $\varphi < \varphi_0$, ce qui constitue une contradiction avec l'hypothèse sur les propriétés de $f(z)$. Ainsi, il existe des m aussi grands que l'on veut et tels que $l_m \geqslant 2^{qm}$; par conséquent, pour un concours de circonstances défavorable, l'erreur commise en arrondissant a_j^m peut dépasser $2^{qm}2^{-t-1}$. D'autre part, on ne saurait espérer une estimation de l'erreur meilleure que (2). En ce qui concerne l'erreur totale résultant de la substitution de $P^m(x)$ à $f(x)$ et des erreurs sur les coefficients du polynôme, la meilleure estimation est bien $\varepsilon(m) = 2^{qm}2^{-t-1} + 2^{-m}$. L'équation du point d'extrémum de la fonction $\varepsilon(m)$ s'écrit:

$$\varepsilon'_m(\bar{m}) = q \ln 2 \cdot 2^{\bar{qm}}2^{-t-1} - \ln 2 \cdot 2^{-\bar{m}} = 0.$$

D'où $2^{\bar{m}} = \left(\frac{2}{q}\right)^{\frac{1}{q+1}} 2^{\frac{t}{q+1}}$; portant cette valeur dans l'expression de $\varepsilon(\bar{m})$, il vient

$$\varepsilon(\bar{m}) = \left(\frac{2}{q}\right)^{\frac{q}{q+1}} 2^{-\frac{t}{q+1}} 2^{-1} + \left(\frac{2}{q}\right)^{-\frac{1}{q+1}} 2^{-\frac{t}{q+1}} = \frac{q+1}{2}\left(\frac{2}{q}\right)^{\frac{q}{q+1}} 2^{-\frac{t}{q+1}}.$$

Donc, en travaillant avec des nombres de longueur t, la fonction ne sera approchée qu'à $\frac{t}{q+1}$ chiffres près au plus. Par exemple, pour la fonction $f(x) = \frac{1}{1+25\,x^2}$, la borne inférieure de φ tels que $f(z)$ soit analytique dans le domaine G_φ est $\varphi^0 = 2 \operatorname{arctg} 0{,}2$; donc $q + 1 \approx \approx 3{,}2$. Une machine d'une capacité de 60 chiffres binaires ne nous permet donc dans le meilleur des cas d'approcher la fonction qu'à 20 chiffres exacts près.

Ces considérations nous incitent à rechercher, pour les polynômes, des formes d'écriture et de calcul plus simples.

Voici un argument en faveur du caractère exceptionnel de cet état de choses. On approche d'ordinaire des fonctions entières ou celles présentant des singularités à une grande distance du cercle unité; les raisonnements ci-dessus ne donnent pas alors lieu à des conclusions tellement pessimistes. Formellement cela est vrai. Cependant, tout en étant entières analytiques, ces fonctions croissent souvent de façon très rapide avec la partie imaginaire de l'argument. Bien que les quantités l_n restent uniformément bornées en n, elles peuvent atteindre des valeurs si grandes que $|l_n| 2^{-t-1}$ est sensiblement supérieur à la précision voulue. Citons à titre d'exemple la fonction $\varphi(x) = \int_0^1 \exp(i\lambda x t^2)\,dt$ pour λ grands.

Considérons les polynômes de meilleure approximation écrits sous forme de somme de valeurs des polynômes de Tchebycheff

$$P_n(x) = \sum_{j=0}^{n} d_j T_j(x); \qquad (9)$$

on a

$$\frac{2}{\pi}\int_{-1}^{1}\frac{T_n(x)\,T_m(x)}{\sqrt{1-x^2}}\,dx=\begin{cases}2 & \text{pour } n=m=0,\\ \delta_{n-m} & \text{pour } n^2+m^2>0.\end{cases}$$

Aussi

$$\frac{2}{\pi}\int_{-1}^{1}\frac{P_n^2(x)}{\sqrt{1-x^2}}\,dx=\frac{2}{\pi}\sum_{j=0}^{n}|d_j|^2\int_{-1}^{1}\frac{T_j^2(x)}{\sqrt{1-x^2}}\,dx=2d_0^2+\sum_{j=1}^{n}d_j^2\geqslant\sum_{j=0}^{n}d_j^2.$$

Utilisons l'inégalité de Cauchy-Bouniakovski

$$\sum_{j=0}^{n}|d_j|=\sum_{j=0}^{n}1\cdot|d_j|\leqslant\sqrt{\left(\sum_{j=0}^{n}1\right)\cdot\left(\sum_{j=0}^{n}d_j^2\right)}=\sqrt{(n+1)\sum_{j=0}^{n}d_j^2}.$$

Il en résulte

$$\sum_{j=0}^{n}|d_j|\leqslant\sqrt{\frac{2}{\pi}(n+1)\int_{-1}^{1}\frac{P_n^2(x)}{\sqrt{1-x^2}}\,dx}.$$

On a pour la suite de polynômes réalisant l'approximation :

$$\frac{2}{\pi}\int_{-1}^{1}\frac{P_n^2(x)}{\sqrt{1-x^2}}\,dx\to\frac{2}{\pi}\int_{-1}^{1}\frac{f^2(x)}{\sqrt{1-x^2}}\,dx.$$

On voit donc que $\sum_{j=0}^{n}|d_j|$ ne croît pas plus vite que $\sqrt{n}$. Puisque $|T_n(x)|\leqslant 1$, l'erreur sur le polynôme, résultat des erreurs sur a_j, ne dépasse pas une quantité de l'ordre de $(\sum_{j=0}^{n}|d_j|)\,2^{-t}=O(\sqrt{n}\,2^{-t})$. En pratique, on ne se sert généralement pas, pour approcher une fonction, des polynômes de degrés très élevés, et cette estimation de l'erreur fait bien l'affaire. Il n'est certes pas obligatoire de représenter les polynômes approximant la fonction par la combinaison linéaire (9) de polynômes de Tchebycheff. Suivant le contexte, on se décidera pour la combinaison linéaire de polynômes d'un autre système orthogonal, par exemple de polynômes de Legendre. Si nous avons préféré les polynômes de Tchebycheff, c'est qu'ils se calculent à l'aide des relations de récurrence particulièrement simples.

Comment obtenir en pratique le polynôme d'approximation sous forme (9)? On peut commencer à rechercher une approximation du polynôme meilleur approximant sous la forme

$$Q_n^k(x)=\sum_{j=0}^{n}d_j^kT_j(x)$$

et à déterminer les coefficients d_j^k à partir du système linéaire d'équations correspondant à (7.2):

$$f(x_i^k) - Q_n^k(x_i^k) = (-1)^i E^k$$

(on fait appel à cet effet à un algorithme stable de résolution des systèmes algébriques linéaires).

Nous avons déjà décrit un deuxième procédé qui aboutit à la forme (1). Ne pensez cependant pas qu'une fois (1) obtenue, on peut écrire le polynôme sous la forme usuelle, puis passer à la représentation (9). Si A_n est grand, le polynôme $P_n(x)$ écrit comme $a_0 + \ldots + a_n x^n$ présente une grande erreur. Celle-ci est hors calculs par rapport aux transformations ultérieures, si bien que (9) la contiendra également. On passe directement de (1) à (9) en sautant l'écriture usuelle: le système de fonctions $T_n(x)$ de poids $2/(\pi\sqrt{1-x^2})$ étant orthogonal, on a

$$a_0 = \frac{1}{\pi}\int_{-1}^{1} \frac{Q_n(x)}{\sqrt{1-x^2}}\,dx, \qquad a_j = \frac{2}{\pi}\int_{-1}^{1} \frac{Q_n(x)\,T_j(x)}{\sqrt{1-x^2}}\,dx.$$

Dans le cas considéré, le degré des polynômes $Q_n T_j(x)$ est au plus $2n$. Aussi la formule de Gauss-Hermite à $(n+1)$ points est-elle exacte pour ces intégrales:

$$a_0 = \frac{1}{n+1}\sum_{q=1}^{n+1} Q_n(y_{n+1}^q), \qquad a_j = \frac{2}{n+1}\sum_{q=1}^{n+1} Q_n(y_{n+1}^q)\,T_j(y_{n+1}^q) \tag{10}$$

pour $j > 0$; ici

$$y_{n+1}^q = \cos\left(\pi\,\frac{2q-1}{2(n+1)}\right),$$

les valeurs nécessaires de $Q_n(y_{n+1}^q)$ étant trouvées par la formule (1).

Remarquons que le processus itératif permet le procédé combiné suivant: les points $x_0^k, \ldots, x_{n+1}^k$ étant donnés, les polynômes $Q_n^k(x)$ sont recherchés par la méthode décrite sous la forme (1). Puis, à l'aide des relations (10), on les met sous la forme (9), en vue d'une recherche plus économe des points x_i^{k+1}.

Résumons la marche à suivre pour calculer les valeurs du polynôme $P_n(x)$ sur $[-1, 1]$ en négligeant le problème de meilleure approximation des fonctions.

Soit $\sum_{j=0}^{n} a_j x^j$ la forme habituelle de ce polynôme et supposons qu'une erreur de l'ordre de $(\sum_{j=0}^{n} |a_j|)\,2^{-t}$ soit admissible. On utilisera alors le schéma de Hörner

$$P_n(x) = a_0 + x\,(a_1 + x\,(\ldots + x\,(a_{n-1} + x a_n)\ldots)).$$

Sinon on obtiendra la représentation

$$P_n(x) = \sum_{j=0}^{n} d_j T_j(x).$$

Pour diminuer le nombre d'opérations effectuées pour trouver $P_n(x)$, on procédera de la façon suivante. Posons

$$D_0 = d_0 - \frac{d_2}{2}, \quad D_j = \frac{d_j - d_{j+2}}{2}$$

pour $j = 1, \ldots, n-2$, $D_{n-1} = \frac{d_{n-1}}{2}$, $v_n = \frac{d_n}{2}$.

En calculant $P_n(x)$, on trouve $y = 2x$, $v_{n-1} = yv_n + D_{n-1}$, puis successivement $v_{n-2}, \ldots, v_0 = P_n(x)$ selon la formule de récurrence $v_k = yv_{k+1} - v_{k+2} + D_k$. L'erreur de calcul de cet algorithme est, on le sait,

$$O\left(n\left(\min\left(n, \frac{1}{\sqrt{1-x^2}}\right)\right)\left(\max_{[-1,\,1]}(P_n(x))\right) 2^{-t}\right).$$

PROBLÈME. Vérifier que $v_0 = P_n(x)$.

§ 9. Procédés de calcul des fonctions élémentaires

On recourt souvent aux polynômes de meilleure approximation pour approcher des fonctions issues de l'expérience ou de calculs compliqués si toute information complémentaire sur les propriétés de ces fonctions fait défaut. En présence d'une telle information, la fonction d'approximation la plus rationnelle peut bien être d'un type différent.

Ainsi, s'agissant des fonctions élémentaires, l'on possède une information abondante sur leurs propriétés analytiques, ce qui permet de réduire considérablement la masse de calcul.

Considérons un programme standard de calcul de $\ln x$. x s'écrit: $x = 2^p X$, où $\frac{1}{2} \leqslant X \leqslant 1$, i.e. on travaille en binaire; par ailleurs $\ln x = p \ln 2 + \ln X$. On calcule $\ln X$ à l'aide de la série

$$\ln X = 2 \sum_{k=0}^{\infty} \frac{1}{2k+1} \left(\frac{X-1}{X+1}\right)^{2k+1}.$$

Si $\frac{1}{2} \leqslant X \leqslant 1$, ses termes décroissent plus vite que 3^{-k}, tandis que la décroissance de la série élémentaire

$$\ln X = \sum_{k=0}^{\infty} \frac{(-1)^{k-1}}{k} (X-1)^k$$

est de l'ordre de $2^{-k}/k$. Pour accélérer la convergence on pose $X = y/\lambda$ et on calcule $\ln y$ en recourant à la série

$$\ln y = 2 \sum_{k=0}^{\infty} \frac{1}{2k+1} \left(\frac{y-1}{y+1}\right)^{2k+1};$$

λ est tel que la quantité $u = \max\limits_{\frac{1}{2}\leqslant X\leqslant 1} \left|\frac{y-1}{y+1}\right|$ soit minimale. Cela donne $\lambda = \sqrt{2}$ et $u < 0{,}172$; la formule de calcul se récrit

$$\ln x = \left(p-\frac{1}{2}\right)\ln 2 + 2v\sum_{k=0}^{\infty}\frac{1}{2k+1}v^{2k},$$

$$v = \frac{\sqrt{2}X-1}{\sqrt{2}X+1}.$$

Afin de réduire le nombre de termes pris en considération dans l'expression $\sum\limits_{k=0}^{\infty}\frac{1}{2k+1}v^{2k}$, cette dernière expression fait place au polynôme de meilleure approximation ou à un polynôme proche sur la partie correspondante de l'intervalle de variation de v. Dans notre cas la formule

$$\ln x \approx (p - 0{,}5)\ln 2 + v\,(2{,}000000815 + 0{,}666445069v^2 + 0{,}415054254v^4)$$

est en erreur de $3\cdot 10^{-8}$ au plus.

Certains constructeurs basent les programmes usuels de calcul des fonctions élémentaires sur la représentation de celles-ci par des fractions continues ou sur leur approximation par des fonctions rationnelles. La fonction tg x peut donc être calculée sous la forme de la fraction continue

$$\operatorname{tg} x = \cfrac{x}{1-\cfrac{x^2}{3-\cfrac{x^2}{5-\ldots}}}$$

Un autre procédé répandu est l'approximation rationnelle de la forme

$$\frac{P_n(x)+Q_m(x)}{P_n(x)-Q_m(x)}.$$

Par exemple, l'approximation

$$\exp x \approx \frac{12(x^2+10)+x(x^2+60)}{12(x^2+10)-x(x^2+60)}$$

garantit une précision de $9\cdot 10^{-9}$ pour $|x| \leqslant \frac{\ln 2}{2}$.

On aurait de la peine à imaginer toute la diversité des méthodes employées lors de la rédaction des programmes usuels de calcul des fonctions élémentaires. Les méthodes sont largement tributaires de la capacité de la machine, du degré de précision imposé, du temps des opérations arithmétiques élémentaires, etc.

Considérons un procédé de ce genre, qui va nous permettre (c'est d'ailleurs la raison pour laquelle nous l'avons préféré aux autres) de discuter une fois de plus le problème de l'erreur de calcul.

Les fonctions élémentaires les plus simples vérifient les relations de récurrence de la forme

$$f(x) = F(f(G(x))) \qquad (1)$$

avec F et G qui se calculent aisément, $0 \leqslant G'(x)$, $0 \leqslant G(x) \leqslant qx$ pour $0 \leqslant$

$\leqslant x \leqslant x_0$, $q < 1$. En répétant l'opération n fois, on a

$$f(x) = F\underbrace{(F(\ldots(F}_{n \text{ fois}}(f(\underbrace{G(G(\ldots(G}_{n \text{ fois}}(x))\ldots). \tag{2}$$

Posons

$$G_n(x) = \underbrace{G(G(\ldots(G}_{n \text{ fois}}(x))\ldots),$$

$$F_n(x) = \underbrace{F(F(\ldots(F}_{n \text{ fois}}(x))\ldots).$$

Grâce à l'écriture (2), l'approximation de $f(x)$ sur le segment $[0, x_0]$ se ramène à son approximation sur le segment partiel $[0, G_n(x_0)]$ avec $G_n(x_0) \leqslant q^n x_0$. Le segment considéré étant petit, on aura une bonne précision pour un approximant de degré sensiblement moindre.

Citons plusieurs relations de la forme (1):

$$\exp x = \left(\exp\frac{x}{2}\right)^2, \quad \cos x = 2\left(\cos\frac{x}{2}\right)^2 - 1,$$

$$\sin x = 2\left(\sin\frac{x}{2}\right)\sqrt{1-\left(\sin\frac{x}{2}\right)^2}, \quad \operatorname{tg} x = \frac{2\operatorname{tg}\frac{x}{2}}{1-\left(\operatorname{tg}\frac{x}{2}\right)^2},$$

$$\operatorname{cotg} x = \frac{1}{2}\left(\operatorname{cotg}\frac{x}{2} - \frac{1}{\operatorname{cotg}\frac{x}{2}}\right), \quad \operatorname{arctg} x = 2\operatorname{arctg}\frac{x}{1+\sqrt{1+x^2}},$$

$$\arcsin x = 2\arcsin\frac{x}{\sqrt{2(1+\sqrt{1-x^2})}}.$$

Celle de $\operatorname{cotg} x$ paraît la plus séduisante vu sa simplicité. Une fois trouvé $u = \operatorname{cotg}\frac{x}{2}$, on calcule

$$\sin x = \frac{2}{u+\frac{1}{u}}, \quad \cos x = \frac{u-\frac{1}{u}}{u+\frac{1}{u}}, \quad \operatorname{tg} x = \frac{2}{u-\frac{1}{u}}.$$

Certaines de ces formules sont assez compliquées, et on ne les utilisera qu'avec circonspection. Les constructions citées déterminent le schéma de calcul suivant. Cherchons une certaine approximation de $f(x)$ sur $[0, G_n(x_0)]$, disons par un polynôme ou un rapport de deux polynômes. Calculons successivement

$$G_1(x) = G(x), \; G_2 = G(G_1(x)), \; \ldots, \; G_n(x) = G(G_{n-1}(x)),$$

la valeur approchée $f^* \approx f(G_n(x))$,

$$F_1(f^*) = F(f^*), \; \ldots, \; F_n(f^*) = F(F_{n-1}(f^*))$$

et posons $f(x) \approx F_n(f^*)$.

PROBLÈME. Supposons qu'on n'arrondit pas et que sur le segment $[0, G_n(x_0)]$ la fonction $f(x)$ est approchée par une série partielle de Taylor de q termes. Démontrer que pour ε petits on peut choisir n et q liés de l'ordre de $O(\sqrt{\ln(1/\varepsilon)})$ et tels que l'erreur sur les valeurs approchées de $f(x)$ sur $[0, x_0]$ ne soit pas supérieure à ε.

Sur l'exemple de la fonction exp x attirons l'attention sur l'éventualité d'une croissance indésirable de l'erreur de calcul lorsqu'on cherche $F_n(f^*)$. Moyennant l'égalité $\exp x = 2^p \exp X$ avec

$$p = \left[\frac{x}{\ln 2}\right], \quad X = \left\{\frac{x}{\ln 2}\right\} \ln 2,$$

le problème se réduit à rechercher les valeurs de exp x sur le segment $[0, \ln 2]$; $\{\}$ désigne la partie fractionnaire. On a

$$\exp X = \left(\exp \frac{X}{2^n}\right)^{2^n} = \left(\ldots\left(\exp\left(\frac{X}{2^n}\right)\right)^2\ldots\right)^2.$$

Quelle que soit la façon dont on calcule $\exp \frac{X}{2^n}$, on les arrondit toujours, et leur erreur peut dépasser 2^{-t-1} (t étant la longueur des nombres manipulés). Au lieu de la valeur exacte $\left(\exp \frac{X}{2^n}\right)^{2^n}$, on arrive à celle approchée $\left(\left(\exp \frac{X}{2^n}\right)+\varepsilon\right)^{2^n}$. Conformément à la formule de Lagrange $y_1^N - y_2^N = N\bar{y}^{N-1}(y_1 - y_2)$, avec $\bar{y}$ compris entre y_1 et y_2. Aussi, dans le cas considéré, on commet une erreur

$$2^n\varepsilon\left(\left(\exp\frac{X}{2^n}\right)+\bar{\varepsilon}\right)^{2^n-1} \text{ avec } |\bar{\varepsilon}| \leqslant \varepsilon.$$

Le dernier facteur est au moins 1, et l'erreur peut donc dépasser 2^{n-t-1}. En général, les programmes usuels doivent garantir une précision des valeurs de la fonction de l'ordre de 2^{-t}. Pour n grands, l'erreur est par conséquent inadmissible.

Dans notre exemple les possibilités d'améliorer la précision sautent aux yeux. Le nombre $u_k = \left(\exp \frac{X}{2^n}\right)^{2^k}$ s'écrit en binaire: $1, \underbrace{0\ldots0}_{n-k \text{ fois}}\ldots$ Ainsi, $n - k + 1$ chiffres sur t servent à mémoriser l'information qu'on connaît déjà. C'est pourquoi, il est peut-être possible d'améliorer la précision des calculs en remplaçant cette information par une information essentielle. Pour ne pas avoir à garder ces caractères binaires connus, introduisons la fonction $v(X) = \exp X - 1$. La relation (1) s'écrit alors $v(X) = v\left(\frac{X}{2}\right)\left(2 + v \times \right.$ $\left. \times \left(\frac{X}{2}\right)\right)$. Les valeurs $v\left(\frac{X}{2^n}\right)$ sont au plus 2^{1-n}. Si elles comportent une erreur de l'ordre de $O(2^{-n-t})$, l'erreur sur $v(X)$, donc sur $\exp X = 1 + v(X)$, sera de l'ordre de $O(n\,2^{-t})$. Nous laissons au lecteur le soin de vérifier ce résultat.

Pour cette même raison la formule de récurrence $\cos x = 2\left(\cos \frac{x}{2}\right)^2 - 1$ n'est pas valable. Au lieu d'utiliser directement cette formule, on posera $u(x) = 1 - \cos x$ et on calculera les valeurs de $u(x)$ par la formule de récurrence $u(x) = 2u\left(\frac{x}{2}\right)\left(2 - u\left(\frac{x}{2}\right)\right)$. S'il s'agit de la fonction $\ln x$, on prendra

$$\ln(1+x) = 2\ln\left(1 + \frac{x}{1+\sqrt{1+x}}\right)$$

au lieu de la relation simple $\ln x = 2 \ln \sqrt{x}$.

On constate la même perte de précision si l'un des nombres $f(0)$ (comme dans le cas de exp x) ou $\lim_{n\to\infty} G_n(x)$ (comme dans le cas de ln x) est fini et non nul.

PROBLÈMES. 1. Effectuer des constructions analogues pour $f(x) = \arccos x$.

2. Vérifier que les relations de récurrence destinées au calcul de

$$\sin x,\ \operatorname{tg} x,\ \operatorname{cotg} x,\ \arcsin x$$

ne nécessitent pas de telles transformations.

3. Supposons que les valeurs des fonctions élémentaires sont calculées sur une machine utilisant la représentation binaire en virgule flottante avec t chiffres, t grand. Supposons encore que les exposants des nombres sur lesquels on travaille sont illimités (de pareilles machines n'existent pas, mais les contraintes sur l'exposant des nombres dans les calculateurs modernes sont telles qu'on néglige souvent la finitude de l'exposant maximal). Démontrer qu'en approchant $f(x)$ sur le segment $[0, G_n(x_0)]$ par une série partielle taylorienne de longueur q pour un choix convenable de n et q (de l'ordre de $\sqrt{t}$) et une écriture adéquate des relations (1), on peut indiquer les algorithmes de calcul des fonctions élémentaires susmentionnées dotés de caractéristiques suivantes: un volume de calculs de l'ordre de $O(\sqrt{t})$, une erreur sur les valeurs approchées de l'ordre de $O(\sqrt{t}/2^t)$ *).

§ 10. Vitesse d'approximation des fonctions de différentes classes

On se propose maintenant d'étudier la vitesse d'approximation par des polynômes et des fonctions rationnelles

$$R_n(x) = \frac{a_0 + \ldots + a_n x^n}{b_0 + \ldots + b_n x^n}$$

selon les propriétés d'analyticité et de dérivabilité des fonctions à approcher.

Il s'agira en l'occurrence d'approcher des fonctions données sur $[-1, 1]$ par des polynômes et des fractions rationnelles et des fonctions périodiques de période 2π par des polynômes et fractions trigonométriques. Soient

$$\|f\|_1 = \sup_{[-1, 1]} |f(x)|,$$

$$\|f\|_2 = \sup_{[0, 2\pi]} |f(x)|$$

les normes respectives.

Les théoriciens ont généralement affaire aux classes types de fonctions suivantes: $C_r(M, [-1, 1])$ (r entier), la classe des fonctions possédant sur $[-1, 1]$ des dérivées $f^{(l)}$, $l = 0, \ldots, r-1$, continues, la dérivée $f^{(r)}$ continue par morceaux et telles que $|f^{(l)}| \leqslant M$ pour $l = 0, \ldots, r$; $H_r(M, 2\pi)$, la classe des fonctions de période 2π dont les dérivées vérifient les mêmes conditions; $A(M, G)$, la classe des fonctions prolongeables analytiquement du segment $[-1, 1]$ dans le domaine ouvert G simplement connexe de la variable complexe et y satisfaisant à la condition $|f(z)| \leqslant M$; $A(M, h, 2\pi)$, la classe des fonctions périodiques de période 2π qui peuvent être prolongées analytiquement dans le domaine $|\operatorname{Im} z| < h$ et vérifiant dans ce domaine la condition $|f(z)| \leqslant M$.

Désignons par E_H les ellipses de foyers points de l'axe réel $[-1, 1]$ et dont les demi-axes forment la somme $H > 1$. Posons $\alpha(G) = \sup_{E_H \in G} H$, ce qui veut dire que $\alpha(G)$ est la somme des demi-axes de la plus grande des ellipses E_H appartenant au domaine fermé $\overline{G}$ (fig. 4.10.1).

*) Une partie des raisonnements de ce paragraphe sont dus à S. Oussov.

Il n'existe qu'une transformation conforme et une seule de points fixes -1 et 1 qui représente G sur l'intérieur d'une ellipse de la famille E_H. Notons $\beta(G)$ la valeur de H correspondant à cette ellipse.

Introduisons les notations suivantes:

$t_n(f)$: erreur dans la meilleure approximation uniforme par les polynômes trigonométriques

$$T_n(x) = a_0 + \sum_{j=1}^{n} (a_j \cos jx + b_j \sin jx),$$

$$t_n = \inf_{a_j,\, b_j} \| f - T_n \|_2 ;$$

$r_n(f)$: erreur dans la meilleure approximation uniforme par les fonctions rationnelles $R_n(x) = \dfrac{a_0 + \ldots + a_n x^n}{1 + b_1 x + \ldots + b_n x^n}$, $r_n(f) = \inf\limits_{a_j,\, b_j} \| f - R_n \|_1$;

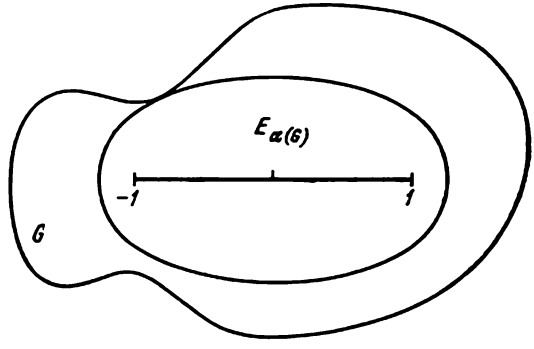


Fig. 4.10.1

$\varepsilon_n(f) = \| f - L_n^f \|_1$, où $L_n^f(x)$ est un polynôme d'interpolation de degré n basé sur les points de Tchebycheff

$$\cos\left(\pi \frac{2m+1}{2n+2}\right), \quad m = 0, \ldots, n ;$$

$\tau_n(f) = \| f - T_n^f \|_2$, où T_n^f est un polynôme d'interpolation trigonométrique de degré n basé sur

$$x_j = \frac{2\pi j}{2n+1}, \quad j = 0, \ldots, 2n ;$$

$\rho_n(f)$: erreur dans la meilleure approximation uniforme par les fractions trigonométriques

$$R_n^T(x) = \frac{a_0 + \sum\limits_{j=1}^{n} (a_j \cos jx + b_j \sin jx)}{1 + \sum\limits_{j=1}^{n} (c_j \cos jx + b_j \sin jx)}.$$

On a les estimations suivantes: pour la classe $C_r(M, [-1, 1])$:

$$E_n(f)=O\left(\frac{1}{n^r}\right),\quad \varepsilon_n(f)=O\left(\frac{\ln n}{n^r}\right),\quad r_n(f)=O\left(\frac{1}{n^r}\right);$$

pour la classe $H_r(M, 2\pi)$:

$$t_n(f)=O\left(\frac{1}{n^r}\right),\quad \tau_n(f)=O\left(\frac{\ln n}{n^r}\right),\quad \rho_n(f)=O\left(\frac{1}{n^r}\right);$$

pour la classe $A(M, G)$:

$$\begin{aligned} E_n(f)&=O(M(\alpha(G))^{-n}),\\ \varepsilon_n(f)&=O(M(\alpha(G))^{-n}),\\ r_n(f)&=O(M(\beta(G))^{-n}); \end{aligned} \tag{1}$$

pour la classe $A(M, h, 2\pi)$:

$$\begin{aligned} t_n(f)&=O(Me^{-2hn}),\\ \tau_n(f)&=O(Me^{-2hn}),\\ \rho_n(f)&=O(Me^{-2hn}). \end{aligned}$$

L'approximation par les fonctions rationnelles trigonométriques

$$R_n^T(x)=\frac{a_0+\sum_{j=1}^{n}(a_j\cos jx+b_j\sin jx)}{1+\sum_{j=1}^{n}(c_j\cos jx+d_j\sin jx)}$$

est intéressante lorsque la fonction périodique est prolongeable analytiquement dans un domaine de la variable complexe autre que la bande $|\operatorname{Im} z|\leqslant h$. La majorité des fonctions qu'on approche sont analytiques; les estimations ci-dessus attestent que pour de telles fonctions la meilleure approximation uniforme ne donne aucun gain sur l'ordre de l'erreur par rapport à l'interpolation utilisant les points de Tchebycheff.

Si le domaine d'analyticité est l'une des ellipses E_H ou, dans le cas périodique, la bande $|\operatorname{Im} z|\leqslant h$, le coût de la meilleure approximation rationnelle est environ deux fois plus élevé tandis que le gain sur l'ordre de l'estimation est nul. Si le domaine d'analyticité est essentiellement autre, l'approximation à l'aide des fractions rationnelles ou des fonctions rationnelles trigonométriques peut donner lieu à un gain.

Le gain est particulièrement important si $\alpha(G)=1$. Soit en effet $f(x)$ une fonction analytique en tous les points du segment $[-1, 1]$, sauf en un nombre fini de points, et satisfaisant à la condition de Hölder avec un $\lambda>0$:

$$|f(x_1)-f(x_2)|\leqslant C|x_1-x_2|^{\lambda}.$$

La fig. 4.10.2 représente un domaine d'analyticité correspondant à ce cas. On dit alors de $E_n(f)$ que

$$E_n(f)=O(n^{-\sigma}),\quad \sigma=\sigma(\lambda)>0.$$

Par ailleurs,

$$r_n(f)=O(e^{-c\sqrt{n}}),\qquad c>0.$$

La meilleure approximation pour des fonctions rationnelles $R_n(x)$ est pratiquement difficile à obtenir, car $R_n(x)$ dépend de façon non linéaire d'un grand nombre de paramètres. On appliquera donc dans certains cas des procédés mixtes tels que ce nombre soit essentiellement moindre. On utilisera parfois

des approximations de la forme

$P_n(P_m(x))$, P_n, P_m polynômes,

$\dfrac{P_n(x)}{(Q_m(x))^l}$, P_n et Q_m polynômes, $n = ml$,

$P_n(R_m(x))$, P_n polynôme et R_m fraction rationnelle.

L'erreur sur la meilleure approximation par de telles fonctions se note respectivement

$$E_{nm}(f),\ r_{nm}(f),\ e_{nm}(f).$$

Prenons le cas d'un domaine G de frontière courbe analytique. Si $\alpha(G) > 1$,

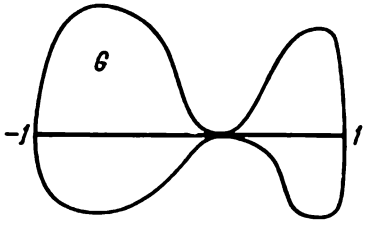


Fig. 4.10.2

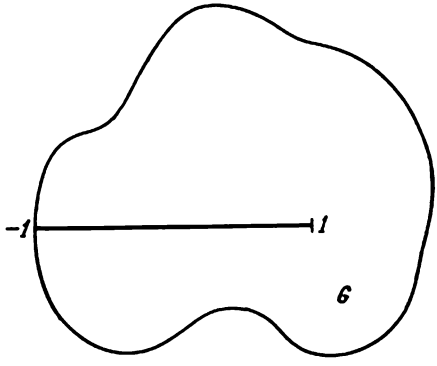


Fig. 4.10.3

on évalue alors de la facon suivante l'erreur pour les fonctions de la classe $A(M, G)$:

$$e_{nm}(f) \leqslant E_{nm}(f) = O(M\beta(G)^{-n}) \qquad \text{pour } m \geqslant \gamma(G) \ln n;$$

et

$$r_{nm}(f) = O(M\beta(G)^{-n}) \qquad \text{pour } m \geqslant \eta(G) \ln n.$$

Si $\alpha(G) = 1$ mais tout point de l'intervalle $(-1, 1)$ est un point d'analyticité (fig. 4.10.3), on a pour $m \geqslant \chi(G) \ln n$ l'estimation

$$r_{nm}(f) = O(M\delta(G)^{-\sqrt{n}}),\ \delta(G) > 1.$$

§ 11. Interpolation et approximation par des fonctions-spline

L'approximation par des fonctions-spline est intéressante en premier lieu par l'appréciation qu'on en donne. Pour les uns, il s'agit d'une méthode universelle de résolution des problèmes de l'analyse numérique qu'on essaye donc dans les domaines les plus divers. Pour d'autres, c'est une mode mathématique passagère. La vérité tient probablement le milieu; en tout cas le champ des applications de cette méthode ne cesse de s'élargir.

Nous parlerons, pour fixer les idées, de l'approximation d'une fonction $f(x)$ sur le segment $[0, 1]$. Partageons ce dernier en parties

$$[x_0, x_1], \ldots, [x_{N-1}, x_N],\ x_0 = 0,\ x_N = 1$$

et désignons cette partition par Δ. On appelle *fonction-spline* d'ordre m, et on la note $S_\Delta^m(f, x)$, une fonction formée de polynômes successifs de degré m

$$S_\Delta^m(f, x) = P_{nm}(x) = a_{n0} + \ldots + a_{nm}x^m \quad \text{pour } x_{n-1} \leqslant x \leqslant x_n \qquad (1)$$

se raccordant jusqu'à la dérivée $(m-1)$-ième comprise aux abscisses $x_1, \ldots, x_{N-1}$:

$$P_{nm}^{(k)}(x_n) = P_{n+1,m}^{(k)}(x_n) \qquad (2)$$

pour $k = 0, \ldots, m-1$, $n = 1, \ldots, N-1$.

On dispose en tout de $Q = N(m+1)$ inconnues a_{nm}, et les relations (2) forment un système de $(N-1)\,m$ équations. Les autres équations des coefficients s'établissent à partir de la condition de proximité de la fonction à approcher et de certaines autres conditions complémentaires.

Considérons le problème élémentaire d'approximation par des fonctions-spline linéaires ($m = 1$). Alors le nombre total Q de paramètres libres est $2N$. Proposons-nous de construire la fonction-spline $S_\Delta^1(f, x)$ coïncidant avec $f(x)$ aux abscisses $x_0, \ldots, x_N$; il vient le système d'équations

$$P_{n1}(x_{n-1}) = f(x_{n-1}), \quad n = 1, \ldots, N,$$
$$P_{n1}(x_n) = f(x_n), \quad n = 1, \ldots, N.$$

Il se décompose en systèmes d'équations relativement aux coefficients des polynômes

$$P_{n1}(x_{n-1}) = a_{n0} + a_{n1}x_{n-1} = f(x_{n-1}),$$
$$P_{n1}(x_n) = a_{n0} + a_{n1}x_n = f(x_n).$$

D'où

$$a_{n1} = (f(x_n) - f(x_{n-1}))/(x_n - x_{n-1}),$$
$$a_{n0} = f(x_{n-1}) - a_{n1}x_{n-1};$$

$P_{n1}(x)$, que nous avons rencontré plus d'une fois, est le polynôme d'interpolation du premier degré basé sur les points x_{n-1}, x_n.

$S_\Delta^1(f, x)$ apparaît également dans le problème suivant. Soit S_1 un ensemble des fonctions continues $s(x)$ satisfaisant aux conditions

$$s(x_n) = f(x_n), \; n = 0, \ldots, N, \qquad I_1(s) = \int_0^1 (s'(x))^2\,dx < \infty.$$

On demande de trouver une fonction $s_1(x) \in S_1$ réalisant

$$\inf_{s \in S_1} I_1(s).$$

On recherche une fonction de norme minimale $\| s' \|_{L_2}$, i.e., dans un certain sens, la fonction la plus régulière de la famille S_1 se confondant avec $f(x)$ en x_n. Pour cette fonctionnelle l'équation d'Euler s'écrit $s''(x) = 0$. Donc, $s_1(x)$ est linéaire sur chacun des segments $[x_{n-1}, x_n]$, et, par conséquent,

$$s_1(x) = S_\Delta^1(f, x).$$

Cette approche de l'approximation d'une fonction, qui engendre tout naturellement la fonction-spline $S_\Delta^1(f, x)$, peut être développée comme suit.

On recherche la fonction d'approximation dans l'espace des fonctions plus régulières. Soit, par exemple, l'ensemble S_2 des fonctions $s(x)$ à la dérivée première continue, qui satisfont aux conditions

$$s(x_n) = f(x_n), \quad n = 0, \ldots, N, \quad I_2(s) = \int_0^1 (s''(x))^2 \, dx < \infty.$$

On demande de trouver une fonction $s_2(x) \in S_2$ réalisant $\inf_{s \in S_2} I_2(s)$. L'équation d'Euler de cette fonctionnelle est $s^{(4)}(x) = 0$, i.e. $s_2(x) = P_{n3}(x) = a_{n0} + a_{n1}x + a_{n2}x^2 + a_{n3}x^3$ sur les segments $[x_{n-1}, x_n]$. Le nombre de paramètres a_{nk} à déterminer vaut $4N$, le nombre de conditions $3N - 1$, dont $2N$ conditions

$$P_{n3}(x_n) = P_{n+1,3}(x_n) = f(x_n) \tag{3}$$

et $N - 1$ conditions de continuité de la dérivée première

$$P'_{n3}(x_n) = P'_{n+1,3}(x_n) \tag{4}$$

aux points $x_1, \ldots, x_{N-1}$. Les conditions qui manquent s'obtiennent comme des conditions aux limites naturelles en x_n.

Afin de vous éviter la peine de consulter le chapitre correspondant du calcul variationnel, déduisons ces conditions de façon indépendante. Supposons que la fonction de minimisation $s_2(x)$ existe et est un polynôme du troisième degré sur les segments $[x_{n-1}, x_n]$. Soit $\eta(x)$ une fonction indéfiniment dérivable:

$$\eta(x_0) = \ldots = \eta(x_N) = 0. \tag{5}$$

Posons

$$F(t) = \int_0^1 (s_2''(x) + t\eta''(x))^2 \, dx =$$

$$= \int_0^1 (s_2''(x))^2 \, dx + 2t \int_0^1 s_2''(x) \, \eta''(x) \, dx + t^2 \int_0^1 (\eta''(x))^2 \, dx.$$

Puisque, par hypothèse, $F(t) \geqslant F(0)$, alors $F'(0) = 2\int_0^1 s_2''(x)\,\eta''(x) \times$ $\times dx = 0$.

Mettons la quantité $B = F'(0)/2$ sous la forme $B = b_1 + \ldots + b_N$,

$$b_n = \int_{x_{n-1}}^{x_n} s_2''(x)\,\eta''(x)\,dx = \int_{x_{n-1}}^{x_n} P_{n3}''(x)\,\eta''(x)\,dx,$$

et intégrons deux fois par parties l'expression de b_n. Il vient

$$b_n = \int_{x_{n-1}}^{x_n} P_{n3}^{(4)}(x)\,\eta(x)\,dx + (P_{n3}''(x)\,\eta'(x) - P_{n3}'''(x)\,\eta(x))\Big|_{x_{n-1}}^{x_n}.$$

Vu (5) et l'égalité $P_{n3}^{(4)}(x) = 0$, on a

$$b_n = P_{n3}''(x)\,\eta'(x)\Big|_{x_{n-1}}^{x_n}.$$

En sommant en n, on trouve

$$B = P_{N3}''(1)\,\eta'(1) + \sum_{n=1}^{N-1} (P_{n3}''(x_n) - P_{n+1,\,3}''(x_n))\,\eta'(x_n) - P_{13}''(0)\,\eta'(0).$$

A vrai dire, nous n'avions pas le droit d'intégrer par parties l'intégrale de départ B, le problème de dérivabilité de $s_2(x)$ aux points $x_1, \ldots, x_{N-1}$ n'ayant pas été tranché.

Puisque $B = 0$ pour $\eta(x)$ quelconque vérifiant (5), il en découle les $N + 1$ équations manquantes

$$P_{13}''(0) = P_{N3}''(1) = 0, \quad P_{n3}''(x_n) - P_{n+1,\,3}''(x_n) = 0, \qquad (6)$$
$$n = 1, \ldots, N-1.$$

Arrêtons-nous sur la résolubilité et la résolution pratique du système d'équations (3), (4), (6). Pour plus de commodité, faisons intervenir les quantités $M_n = s_2''(x_n)$. Puisque $s_2''(x)$ est linéaire sur $[x_{n-1}, x_n]$, on a

$$s_2''(x) = M_{n-1}\frac{x_n - x}{h_n} + M_n\frac{x - x_{n-1}}{h_n} \text{ sur } [x_{n-1}, x_n];$$

ici $h_n = x_n - x_{n-1}$. Ces relations et les conditions

$$s_2(x_{n-1}) = f(x_{n-1}),\ s_2(x_n) = f(x_n)$$

nous fournissent

$$s_2(x) = P_{n3}(x) = M_{n-1}\frac{(x_n - x)^3}{6h_n} + M_n\frac{(x - x_{n-1})^3}{6h_n} +$$
$$+ \left(f(x_{n-1}) - \frac{M_{n-1}h_n^2}{6}\right)\frac{x_n - x}{h_n} +$$
$$+ \left(f(x_n) - \frac{M_n h_n^2}{6}\right)\frac{x - x_{n-1}}{h_n} \text{ sur } [x_{n-1}, x_n]. \qquad (7)$$

Les conditions

$$P'_{n3}(x_n) = P'_{n+1,3}(x_n),\ n = 1, \ldots, N-1,$$

constituent les équations

$$\frac{h_n}{6} M_{n-1} + \frac{h_n + h_{n+1}}{3} M_n + \frac{h_{n+1}}{6} M_{n+1} = \\ = \frac{f(x_{n+1}) - f(x_n)}{h_{n+1}} - \frac{f(x_n) - f(x_{n-1})}{h_n}. \quad (8)$$

On a de plus

$$P''_{13}(x_0) = 0, \quad P''_{N3}(x_N) = 0,$$

i.e. $M_0 = 0$, $M_N = 0$. Par la substitution de $M_0 = 0$ et $M_N = 0$ respectivement dans la première et la dernière équation (7) on arrive au système

$$C\mathbf{M} = \mathbf{d} \quad (9)$$

de $N-1$ équations à $N-1$ inconnues:

$$\mathbf{M} = (M_1, \ldots, M_{N-1})^T,\ \mathbf{d} = (d_1, \ldots, d_{N-1})^T.$$

Conformément à (8), les éléments $c_{i,j}$, $i, j = 1, \ldots, N-1$, de la matrice C sont donnés par

$$c_{ij} = \begin{cases} \frac{h_i}{6} & \text{pour } j = i-1, \\ \frac{h_i + h_{i+1}}{3} & \text{pour } j = i, \\ \frac{h_{i+1}}{6} & \text{pour } j = i+1, \\ 0 & \text{pour } |i-j| > 1, \end{cases}$$

et les éléments d_i de la colonne $\mathbf{d}$ par les relations

$$d_i = \frac{f(x_{i+1}) - f(x_i)}{h_{i+1}} - \frac{f(x_i) - f(x_{i-1})}{h_i}.$$

On montrera au ch. IX que ce système se résout par la méthode du balayage au bout de $O(N)$ opérations arithmétiques. Après avoir trouvé M_j, déterminons les polynômes $P_{n3}(x)$ par la formule (7).

PROBLÈMES. 1. Démontrer que la solution du système (9) vérifie l'inégalité

$$\max_{1 \leqslant n \leqslant N-1} |M_n| \leqslant \frac{3}{\min\limits_{1 \leqslant n \leqslant N} h_n} \max_{1 \leqslant n \leqslant N-1} |d_n|;$$

ceci étant, montrer que le système d'équations (3), (4), (6) possède une solution unique.

2. Soit $S^2_\Delta(f, x)$ une fonction-spline associée à la solution du système (9). Démontrer que

$$I_2(f) = I_2(S^2_\Delta) + I_2(f - S^2_\Delta).$$

Cette proposition implique l'unicité de la fonction $s_2(x)$ réalisant le minimum.

Ici la fonction-spline obtenue coïncide avec $f(x)$ en tous les points, ce qui incite à donner à de telles fonctions le nom de fonctions-spline *d'interpolation.*

On dispose souvent d'une information supplémentaire sur le comportement de $f(x)$ aux points $x = 0$, $x = 1$. Il se peut que l'on ait à résoudre par exemple une équation différentielle avec les conditions aux limites

$$f'(0) + \alpha f(0) = \beta, \quad f'(1) + \gamma f(1) = \delta.$$

Celles-ci remplacent alors les conditions aux limites naturelles donnant lieu aux équations $M_0 = M_N = 0$.

Soit le cas où les valeurs de $f(x_n)$ sont données avec une certaine erreur. Il existe toute une série de procédés naturels permettant d'approcher de telles fonctions génératrices de fonctions-spline.

a) Supposons que les valeurs données f_n et les valeurs exactes $f(x_n)$ vérifient

$$|f_n - f(x_n)| \leqslant \delta.$$

Le problème peut s'énoncer comme suit : trouver une fonction $s(x)$ continûment

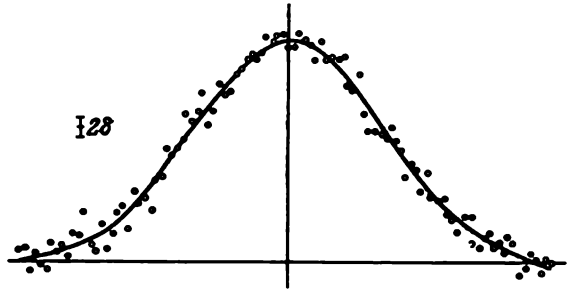


Fig. 4.11.1

dérivable avec $\int_0^1 (s'')^2\,dx < \infty$ qui réalise $\inf I_2(s)$ sur la classe de fonctions satisfaisant à $|s(x_n) - f_n| \leqslant \delta$. Le coût de ce problème est assez élevé. Aussi préfère-t-on parfois rechercher $\inf I_2(s)$ parmi des fonctions telles que

$$\sum_{n=0}^{N} (s(x_n) - f_n)^2 \leqslant c(N+1)\delta^2. \tag{10}$$

b) Les différences $f_n - f(x_n)$ sont, on le sait, des quantités aléatoires de variance δ. Dans ce cas, on minimise également $I_2(s)$ dans la condition (10) ; le choix de $c = O(1)$ se fait à partir du niveau de confiance donné.

Il arrive souvent que les deux membres de (10) coïncident au point de minimum de $I_2(s)$. C'est pourquoi, au lieu du problème de départ, on réalise

le minimum libre en λ et s de la fonctionnelle

$$J_2(\lambda, s) = I_2(s) + \lambda \left(\sum_{n=0}^{N} (s(x_n) - f_n)^2 - c(N+1)\delta^2 \right).$$

La fig. 4.11.1 correspond à un cas concret d'approximation pour $N = 100$ et $c = 0{,}9$. Pour tout λ fixe, la quantité

$$J(\lambda) = \inf_{s} J_2(\lambda, s)$$

est déterminée en résolvant un système d'équations linéaires. On trouve $\inf_{\lambda} J(\lambda)$ par une méthode de minimisation des fonctions d'une variable (cf. ch. VII).

L'expérience accumulée (ou tout simplement le bon sens) peut nous suggérer une valeur approchée de λ correspondant au point d'extrémum libre, $\lambda \approx \lambda_0$; on se borne alors à minimiser la fonctionnelle $J_2(\lambda_0, s)$.

c) Un dernier procédé d'approximation par des fonctions-spline. Soient f_j les valeurs approchées de $f(z_j)$, $j = 1, \ldots, M$. Une information a priori sur la fonction à approcher nous autorise à diviser le segment $[0, 1]$ en $[x_0, x_1], \ldots, [x_{N-1}, x_N]$ avec $N \leqslant M$. L'approximation s'obtient sous forme de la fonction-spline $S_{\Delta}^{m}(f, x)$ minimisant la fonctionnelle

$$\sum_{j=1}^{M} p_j (s(x_j) - f(x_j))^2.$$

§ 12. Entropie et ε-entropie

On dira de façon conventionnelle que tout être est donné par deux sortes d'information :

1) l'information sur un certain ensemble auquel cet être appartient et

2) l'information qui repère l'être en question parmi les éléments de l'ensemble.

Ainsi, tel polynôme de degré m est donné par 1) l'information qui dit que c'est bien un polynôme de degré m, 2) le tableau de ses coefficients.

Notre but est de discuter la quantité d'information minimale permettant de se donner les éléments d'un ensemble. Dans le cas d'approximation de fonctions, la tradition et la commodité des calculs font en général opter pour la donnée d'une fonction à l'aide du tableau de ses valeurs. Cependant la quantité d'information minimale est un problème qui doit être abordé sous une optique plus générale.

Supposons que l'ensemble G dont on veut se donner les éléments comprend N éléments $g_1, \ldots, g_N$ distincts et que ceux-ci sont donnés par un jeu de n caractères binaires $y_1, \ldots, y_n = 0, 1$. Pour pouvoir se donner de façon univoque $g_k \in G$, il faut que N ne soit pas supérieur au nombre de ces jeux. Ils sont 2^n en tout, et la condition nécessaire d'univocité est donc $N \leqslant 2^n$. D'autre part, pour $N \leqslant 2^n$ on peut faire correspondre à tout jeu $y_1, \ldots, y_n$

un nombre naturel

$$j = 1 + \sum_{i=1}^{n} 2^{n-i} y_i. \tag{1}$$

Si $y_1, \ldots, y_n$ varient, le nombre j parcourt toutes les valeurs de 1 à 2^n ; donnons-nous l'élément g_j par le jeu $y_1, \ldots, y_n$ définissant j selon (1). $\log_2 N$ est appelée *entropie* $H(G)$ de G. Désignons par $J(G)$ le nombre minimal de caractères binaires nécessaires pour déterminer univoquement les éléments de G et par $\bar{y}$ l'entier le plus petit non inférieur à y. Nous avons vu que ce nombre minimal de caractères binaires doit être égal à l'entier le plus petit n tel que $N \leqslant 2^n$, par conséquent

$$J(G) = \overline{\log_2 N} = \overline{H(G)}. \tag{2}$$

PROBLÈMES. 1. Démontrer les inégalités

$$J\left(\bigcup_{j=1}^{q} G_j\right) \leqslant \overline{\log_2 q} + \max_{1 \leqslant j \leqslant q} J(G_j), \tag{3}$$

$$J(G_1 \times \ldots \times G_q) \leqslant \sum_{j=1}^{q} J(G_j), \tag{4}$$

$\times$ est le signe du produit direct, i.e. les éléments $G_1 \times \ldots \times G_q$ sont les jeux $(g^1, \ldots, g^q)$ pour $g^i \in G_i$.

2. Donner des exemples concrets d'égalités ou d'inégalités strictes dans (3), (4).

3. On propose de choisir mentalement un entier compris entre 0 et 1000. Pour le deviner, on pose des questions auxquelles on ne répond que par oui ou non. Démontrer qu'au bout de dix questions on devine ce nombre, et au bout de neuf non.

4. On a N boîtes semblables dont $N - 1$ sont de même poids et une est plus lourde. On autorise n pesées sur une balance sans poids. Démontrer qu'avec $N \leqslant 3^n$ il y a garantie de repérer la boîte la plus lourde et qu'avec $N > 3^n$ on n'y arrive pas.

5. Il y a toujours N boîtes semblables mais toutes de poids différent. Evaluer l'ordre de grandeur du nombre minimal n de pesées sur une balance sans poids tel qu'il soit possible d'ordonner les boîtes suivant leur poids.

RÉPONSE. $n \sim \log_2 N!$ pour $N \to \infty$.

Soit maintenant G un ensemble composé d'une infinité d'éléments. En particulier, conformément à (2), tout nombre fini fixe de caractères binaires est insuffisant pour donner de façon univoque les éléments de G.

On pourrait cependant parler de la quantité d'information minimale dont il faut disposer pour donner à ε près les éléments de G dans la métrique d'un espace métrique F, $G \in F$. Voici le procédé pour se donner les éléments $g \in G$. g est déterminé par un certain jeu de caractères binaires $y_1, \ldots, y_n$; on fait correspondre au jeu $y_1, \ldots, y_n$ un élément $f_{y_1, \ldots, y_n} \in F$ tel que $\rho(f_{y_1, \ldots, y_n}, g) \leqslant \varepsilon$. Notons $J_\varepsilon(G; F)$ le nombre minimal de caractères binaires qui

détermine, dans cette méthode, les éléments de G. Soit Φ une extension métrique de F, i.e. un espace métrique contenant F, telle que pour tout couple d'éléments $f_1, f_2 \in F$ on ait

$$\rho_F(f_1, f_2) = \rho_\Phi(f_1, f_2);$$

l'indice inférieur désigne l'espace muni de la métrique en question. Du moment que tout procédé d'approximation de g par les éléments de F peut jouer pour approcher g par les éléments de Φ, on a

$$J_\varepsilon(G; \Phi) \leqslant J_\varepsilon(G; F). \tag{5}$$

Dans la suite l'espacè F sera supposé fixe et l'indice correspondant des notations sera parfois omis.

LEMME 1. *Supposons que les éléments* $f_1, \ldots, f_N \in F$ *forment un* ε*-réseau pour l'ensemble* G. *Alors*

$$J_\varepsilon(G; F) \leqslant \overline{\log_2 N}.$$

DÉMONSTRATION. Selon le lemme, on peut faire correspondre à tout $g \in G$ un certain $f_j \in F$ pour lequel $\rho(g, f_j) \leqslant \varepsilon$. Soit $y_1, \ldots, y_n$ un jeu de symboles binaires correspondant à j d'après (1). En se donnant le jeu $y_1, \ldots, y_n$ on détermine par là même l'élément f_j tel que $\rho(g, f_j) \leqslant \varepsilon$.

LEMME 2. *Soit* $G' \subset G$. *Alors* $J_\varepsilon(G'; \Phi) \leqslant J_\varepsilon(G; \Phi)$ *quel que soit* Φ.

La démonstration est immédiate.

G' est dit η-*discernable* si la distance de deux quelconques de ses éléments est supérieure à η.

LEMME 3. *Soit* $G' \subset G$ *avec* G' 2ε*-discernable et composé de* N *éléments. Alors*

$$\overline{\log_2 N} \leqslant J_\varepsilon(G; \Phi),$$

où Φ *est une extension arbitraire de* F.

DÉMONSTRATION. Il suffit de montrer que

$$\overline{\log_2 N} \leqslant J_\varepsilon(G'; \Phi). \tag{6}$$

Utilisant le schéma du lemme 2 on obtient l'inégalité cherchée. Selon la définition du procédé de détermination des fonctions g l'inégalité $J_\varepsilon(G'; \Phi) \leqslant n$ s'interprète comme suit. Il existe $M \leqslant 2^n$ éléments $\varphi_1, \ldots, \varphi_M \in \Phi$ vérifiant la condition: il correspond à tout $g \in G'$ un certain φ_j tel que $\rho_\Phi(g, \varphi_j) \leqslant \varepsilon$. Un même φ_j ne peut correspondre à des $g_i, g_k \in G'$ distincts. Un raisonnement par l'absurde nous conduit à une contradiction vu les inégalités

$$2\varepsilon < \rho_F(g_i, g_k) = \rho_\Phi(g_i, g_k) \leqslant \rho_\Phi(g_i, \varphi_j) + \rho_\Phi(g_k, \varphi_j) \leqslant 2\varepsilon.$$

Ainsi, le nombre N d'éléments de G' est au plus M, i.e.

$$N \leqslant M \leqslant 2^n = 2^{J_\varepsilon(G';\,\Phi)}.$$

Nous avons obtenu (6) et démontré donc le lemme 3.

Faisons intervenir la quantité

$$J_\varepsilon(G) = \inf_{\Phi} J_\varepsilon(G;\,\Phi);$$

la borne inférieure est prise sur l'ensemble de toutes les extensions Φ de l'espace métrique F.

En bornant inférieurement sur Φ, compte tenu des lemmes 1-3, on obtient respectivement

1) $$J_\varepsilon(G) \leqslant \log_2 N,$$

2) $$J_\varepsilon(G') \leqslant J_\varepsilon(G),$$

3) $$\overline{\log_2 N} \leqslant J_\varepsilon(G).$$

Les lemmes et leurs conséquences donnent la possibilité de majorer et minorer $J_\varepsilon(G;\,F)$.

On appelle *ε-entropie* $H_\varepsilon(G)$ de l'ensemble G la quantité $\log_2 N_\varepsilon(G)$, où $N_\varepsilon(G)$ est le nombre minimal d'ensembles de diamètre ε recouvrant G. Si G est compact, on sait que

$$H_\varepsilon(G) \leqslant J_\varepsilon(G) \leqslant [H_\varepsilon(G)] + 1,$$

la borne inférieure étant prise sur tous les Φ possibles.

Examinons les exemples.

1. F est l'axe réel et G l'ensemble de points

$$x \in [0,\,1],\ \rho(x_1,\,x_2) = |x_1 - x_2|.$$

Posons $N_\varepsilon = \left(\overline{\frac{1}{2\varepsilon}}\right)$. Les points $\varepsilon,\ 3\varepsilon,\ \ldots,\ (2N_\varepsilon - 1)\varepsilon$ forment un ε-réseau. Aussi

$$J_\varepsilon(G;\,F) \leqslant \overline{\log_2 N_\varepsilon}.$$

Selon la définition de N_ε, on a $(N_\varepsilon - 1)\,2\varepsilon < 1$, donc $\varepsilon' = \frac{1}{2}\,(N_\varepsilon - 1)^{-1} > \varepsilon$. N_ε points $0,\ 2\varepsilon',\ \ldots,\ 1$ forment un ensemble 2ε-discernable et, conformément au lemme 3,

$$J_\varepsilon(G;\,\Phi) \geqslant \overline{\log_2 N_\varepsilon}.$$

On a en définitive

$$J_\varepsilon(G) = J_\varepsilon(G;\,F) = \overline{\log_2\left(\overline{\frac{1}{2\varepsilon}}\right)}.$$

2. Problème. Soit F un espace euclidien s-dimensionnel de points $P(x_1,\ \ldots,\ x_s)$, $\rho(P_1,\,P_2)$ égal à une norme du vecteur

P_1P_2 et G un ensemble de points appartenant à la boule $\rho(P, 0) \leqslant A_2 < \infty$ et contenant la boule $\rho(P, 0) \leqslant A_1 \neq 0$. Démontrer l'égalité asymptotique

$$J_\varepsilon(G) = s \log_2 \frac{1}{\varepsilon} + O(1).$$

3. Soit G un ensemble de fonctions $g(x)$ continues, définies sur $[0, 1]$ et telles que $g(0) = 0$ et $|g(x_2) - g(x_1)| \leqslant |x_2 - x_1|$. On pose

$$\rho(f_1, f_2) = \sup_{[0, 1]} |f_1(t) - f_2(t)|.$$

L'espace le plus large pouvant contenir les approximations des fonctions de la classe considérée est l'espace de toutes les fonctions définies sur $[0, 1]$. Désignons-le par F. Posons $n = \overline{\left(\frac{1}{\varepsilon}\right)} - 1$, ce qui signifie $n\varepsilon < 1 \leqslant (n+1)\varepsilon$.

Démontrons que $J_\varepsilon(G; F) = n$. D'abord nous obtenons l'estimation $J_\varepsilon(G; F) \geqslant n$. Il est immédiat que $\varepsilon' = \frac{1}{n} > \varepsilon$; faisons corespondre à tout jeu de caractères binaires $y_1, \ldots, y_n$ une fonction continue $g_{y_1, \ldots, y_n}(x)$ définie par les relations

1) $g_{y_1, \ldots, y_n}(0) = 0$,

2) sur le segment $[(j-1)\varepsilon', j\varepsilon']$, la fonction $g_{y_1, \ldots, y_n}(x)$ est linéaire et

$$(g_{y_1, \ldots, y_n}(x))' = \begin{cases} 1 \text{ si } y_j = 1, \\ -1 \text{ si } y_j = 0. \end{cases} \tag{7}$$

Puisque la dérivée $(g_{y_1, \ldots, y_n}(x))'$ est continue par morceaux on a

$$g_{y_1, \ldots, y_n}(x_2) - g_{y_1, \ldots, y_n}(x_1) = \int_{x_1}^{x_2} (g_{y_1, \ldots, y_n}(x))' \, dx;$$

compte tenu de (7), il vient

$$|g_{y_1, \ldots, y_n}(x_2) - g_{y_1, \ldots, y_n}(x_1)| \leqslant |x_2 - x_1|.$$

Ainsi, la fonction $g_{y_1, \ldots, y_n}(x)$ appartient à la classe considérée. Prenons deux jeux distincts

$$(y_1^1, \ldots, y_n^1) \quad \text{et} \quad (y_1^2, \ldots, y_n^2).$$

Il existe évidemment un $j < n$ tel que $y_i^1 = y_i^2$ si $i < j$ tandis que $y_j^1 \neq y_j^2$. Les fonctions $g_{y_1^l, \ldots, y_n^l}(x)$, $l = 1, 2$, s'annulent dès que $x = 0$; si $0 \leqslant x \leqslant (j-1)\varepsilon'$, leurs dérivées coïncident en vertu de (7). Elles coïncident donc partout avec cette condition, en particulier, au point $(j-1)\varepsilon'$. Sur le segment $[(j-1)\varepsilon', j\varepsilon']$, selon (7), la dérivée de la différence de ces fonctions est constante et égale à 2 en valeur absolue. Aussi

$$g_{y_1^1, \ldots, y_n^1}(j\varepsilon') - g_{y_1^2, \ldots, y_n^2}(j\varepsilon') = \int_{(j-1)\varepsilon'}^{j\varepsilon'} (\pm 2) \, dx = \pm 2\varepsilon'.$$

Ce qui veut dire

$$\rho\left(g_{y_1^1, \ldots, y_n^1}, g_{y_1^2, \ldots, y_n^2}\right) \geqslant 2\varepsilon' > 2\varepsilon.$$

Nous avons donc 2^n fonctions 2ε-discernables. Par conséquent

$$n \leqslant J_\varepsilon(G; F). \tag{8}$$

Passons à la majoration. Posons $\varepsilon'' = \frac{1}{n+1}$ et considérons toutes les fonctions continues $f_{y_1, \ldots, y_n}(x)$ $(y_1, \ldots, y_n = 0, 1)$, définies par

$$f_{y_1, \ldots, y_n}(x) = 0, \quad 0 \leqslant x \leqslant \varepsilon'';$$

sur les segments $[(j-1)\varepsilon'', j\varepsilon'']$, la fonction est linéaire et

$$(f_{y_1, \ldots, y_n}(x))' = \begin{cases} 1 \text{ si } y_{j-1} = 1, \\ -1 \text{ si } y_{j-1} = 0. \end{cases}$$

Puisque $\varepsilon'' + n\varepsilon'' = 1$, ces relations déterminent cette fonction sur tout le segment $[0, 1]$. Montrons que l'ensemble de ces fonctions forme un ε''-réseau. A cette fin, fixons pour tout $g(x) \in G$ une certaine fonction $f_{y_1, \ldots, y_n}(x)$ satisfaisant à $\rho(g, f_{y_1, \ldots, y_n}) \leqslant \varepsilon''$. Déterminons de proche en proche les valeurs $y_1, \ldots, y_n$ correspondantes: notons que sur $[0, j\varepsilon'']$ les valeurs de $f_{y_1, \ldots, y_n}(x)$ ne sont définies que par $y_1, \ldots, y_{j-1}$. Supposons connues $y_1, \ldots, y_{j-1}$ telles que

$$|g(x) - f_{y_1, \ldots, y_n}(x)| \leqslant \varepsilon'' \tag{9}$$

pour $0 \leqslant x \leqslant j\varepsilon$. Il est inutile d'indiquer de telles valeurs pour $j = 1$ car pour toutes les fonctions de la classe considérée sur $[0, \varepsilon'']$ on a

$$|g(x) - f_{y_1, \ldots, y_n}(x)| = |g(x)| = |g(x) - g(0)| \leqslant x \leqslant \varepsilon''.$$

Ainsi, l'hypothèse de départ pour la démonstration par récurrence se trouve vérifiée.

Démontrons que pour le choix

$$y_j = \begin{cases} 1 \text{ si } g((j+1)\varepsilon'') \geqslant f_{y_1, \ldots, y_{j-1}, \ldots, y_n}(j\varepsilon''), \\ 0 \text{ si } g((j+1)\varepsilon'') < f_{y_1, \ldots, y_{j-1}, \ldots, y_n}(j\varepsilon''), \end{cases}$$

la relation (9) est également valable sur le segment $[j\varepsilon'', (j+1)\varepsilon'']$.

Considérons le cas $y_j = 1$ et écrivons

$$|g(x) - g((j+1)\varepsilon'')| \leqslant |x - (j+1)\varepsilon''|.$$

D'où

$$g(x) \geqslant g((j+1)\varepsilon'') + (x - (j+1)\varepsilon'') \geqslant \{f_{y_1, \ldots, y_{j-1}, \ldots, y_n}(j\varepsilon'') + (x - j\varepsilon'')\} - \varepsilon''$$

pour $j\varepsilon'' \leqslant x \leqslant (j+1)\varepsilon''$. Par définition l'expression entre accolades est bien $f_{y_1, \ldots, y_j, \ldots, y_n}(x)$; on trouve donc

$$g(x) \geqslant f_{y_1, \ldots, y_j, \ldots, y_n}(x) - \varepsilon'' \text{ pour } j\varepsilon'' \leqslant x \leqslant (j+1)\varepsilon''.$$

D'autre part,

$$g(x) \leqslant g(j\varepsilon'') + (x - j\varepsilon''),$$

et en substituant $x = j\varepsilon''$ dans (9), il vient

$$|f_{y_1, \ldots, y_{j-1}, \ldots, y_n}(j\varepsilon'') - g(j\varepsilon'')| \leqslant \varepsilon'',$$

donc

$$g(x) \leqslant (f_{y_1, \ldots, y_{j-1}, \ldots, y_n}(j\varepsilon'') + (x - j\varepsilon'')) + \varepsilon''$$

et

$$g(x) \leqslant f_{y_1, \ldots, y_j, \ldots, y_n}(x) + \varepsilon'' \quad \text{pour} \quad j\varepsilon'' \leqslant x \leqslant (j+1)\varepsilon'',$$

ce qui donne enfin

$$|g(x) - f_{y_1, \ldots, y_j, \ldots, y_n}(x)| \leqslant \varepsilon'' \quad \text{sur} \quad [j\varepsilon'', (j+1)\varepsilon''].$$

Le cas $y_j = 0$ se traite de façon analogue. Nous avons ainsi fait un pas de plus dans notre construction de la fonction $f_{y_1, \ldots, y_n}(x)$. En procédant de cette manière on peut arriver jusqu'à l'extrémité du segment.

Vu que $\varepsilon'' = \dfrac{1}{n+1} \leqslant \varepsilon$, l'ensemble de 2^n fonctions que nous venons de construire constitue un ε-réseau. Ces fonctions appartiennentle à G, ensemb de départ. Le lemme 1 fournit donc

$$J_\varepsilon(G; F) \leqslant n. \tag{10}$$

En réunissant cette estimation avec (8) on trouve

$$J_\varepsilon(G; F) = n = \frac{1}{\varepsilon} - 1.$$

N'oublions pas que les éléments du ε-réseau obtenu ci-dessus sont ceux de G.

PROBLÈMES. 1. Démontrer, en se rappelant cette remarque, que dans le cas considéré

$$J_\varepsilon(G) = n.$$

2. Soit G un espace de fonctions à la dérivée $(r-1)$-ième continue, déterminées sur $[0, a]$ et vérifiant

$$f(0) = \ldots = f^{(r-1)}(0) = 0, \quad |f^{(r-1)}(x_2) - f^{(r-1)}(x_1)| \leqslant$$
$$\leqslant A\,|x_2 - x_1| \text{ pour } x_1, x_2 \in [0, a];$$

supposons que F et la métrique sont ceux de l'exemple précédent. $J_\varepsilon(G; F)$ est fonction des paramètres ε, A, a. Démontrer qu'elle est également fonction de la combinaison sans dimension $a\sqrt[r]{A/\varepsilon}$, i.e.

$$J_\varepsilon(G; F) = \psi\left(a\sqrt[r]{A/\varepsilon}\right).$$

3. Démontrer

$$\ln \psi(y) = \ln y + O(1)$$

pour $y \to \infty$.

4. Soit G un espace de fonctions analytiques $A(M, \Omega)$,

$$\rho(f, g) = \sup_{[-1, 1]} |f - g|,$$

$\alpha(\Omega) > 1$ (voir § 10 pour la définition de $A(M, \Omega)$). Démontrer

$$\log_2 H_\varepsilon(G) \sim 2 \log_2 \log_2 \frac{1}{\varepsilon}.$$

Bibliographie du chapitre IV

§§ 3, 4: II [38], §§ 5-7: II [26], § 10: II [7, 45], III [22]; § 11: II [30], III [63]; § 12: II [3, 33], III [31].

CHAPITRE V

PROBLÈMES À PLUSIEURS DIMENSIONS

Comme nous l'avons vu, les problèmes d'approximation, d'interpolation d'intégration et de dérivation numériques des fonctions d'une variable sont étudiés en détail, si bien qu'on dispose à l'heure actuelle de systèmes assez développés de programmes standard de résolution des problèmes à une variable. La plupart des résultats de recherches portant sur une variable s'étendent sans difficulté au cas de deux ou plusieurs variables, mais certaines méthodes qui apparaissent en l'occurrence sont souvent peu constructives.

Au cours des constructions théoriques on divise les problèmes en classes, par exemple, la classe des problèmes de calcul d'intégrales des fonctions aux dérivées bornées, puis on procède à des études relatives à ces classes. La description adoptée ne caractérise certes pas toujours (loin de là!) de façon satisfaisante telle classe de problèmes réels, mais la vitesse des calculateurs électroniques modernes est telle qu'en dépit de la description très sommaire des problèmes unidimensionnels on réussit à résoudre la plupart d'entre eux en utilisant des méthodes standard issues des études théoriques.

Dans le cas de deux ou plusieurs variables la description adoptée n'est pas non plus toujours réussie. Le coût des problèmes s'accroît brusquement avec leur dimension, c'est pourquoi on n'arrive généralement pas à mettre au point des méthodes standard de résolution de vastes classes de problèmes à plusieurs dimensions qui soient aussi précises que dans le cas unidimensionnel.

D'autre part, les problèmes mathématiques à plusieurs dimensions sont généralement issus de la description de processus compliqués. Ces descriptions étant habituellement grossières, la précision exigée est plus faible que dans le cas unidimensionnel. Par exemple, la précision demandée des solutions des équations de la dynamique des gaz est sensiblement plus basse qu'en balistique ou en mécanique céleste.

Ceci ne doit pas inciter à penser que la résolution des problèmes à plusieurs dimensions est une tâche quasi impossible. La différence fondamentale d'avec le cas unidimensionnel est probablement leur caractère très diversifié, ce qui exige la coopération de spécialistes plus qualifiés. L'étude de classes restreintes de problèmes multidimensionnels et une plus grande rapidité des ordinateurs entraîneront

la création des méthodes standard de résolution des problèmes de certaines classes et le concours d'opérateurs de qualification plus modeste.

§ 1. Méthode des coefficients indéterminés

La résolution de nombreux problèmes à plusieurs dimensions se réduit souvent à celle des problèmes élémentaires suivants.

Soient donnés, dans un domaine d'un espace G de dimension s, les points $P_1, \ldots, P_N$ et les valeurs d'une fonction f en ces points. On désire

1) obtenir une valeur approchée de la fonction $f(P)$;
2) obtenir une valeur approchée d'une dérivée $\mathcal{D}f$ au point P;
3) calculer l'intégrale

$$I(f)=\int_G f(P)\,p(P)\,dP,$$

$p(P)$ étant la fonction de poids.

Le procédé le plus simple est en l'occurrence la *méthode des coefficients indéterminés*. Admettons qu'on connaît, par des considérations quelconques, que la fonction $f(P)$ est bien approchée par les combinaisons linéaires

$$\sum_{j=1}^{N} B_j\omega_j(P).$$

Exigeons qu'une telle combinaison linéaire se confonde avec $f(P)$ aux points donnés:

$$\sum_{j=1}^{N} B_j\omega_j(P_q)=f(P_q), \quad q=1,\ldots,N. \tag{1}$$

Supposons que $\det \| \omega_j(P_q) \| \neq 0$; alors la matrice $\| \omega_j(P_q) \|$ admet pour inverse $A = \| a_{jq} \|$ et la solution (1) est de la forme suivante:

$$B_j=\sum_{q=1}^{N} a_{jq}f(P_q). \tag{2}$$

La fonction

$$g(P)=\sum_{j=1}^{N} B_j\omega_j(P)$$

se confond avec $f(P)$ aux points P_q. La substitution de B_j de (2) donne une autre représentation de $g(P)$:

$$g(P)=\sum_{q=1}^{N} z_q(P)\,f(P_q), \quad \text{où} \quad z_q(P)=\sum_{j=1}^{N} a_{jq}\omega_j(P). \tag{3}$$

Cette écriture de la fonction d'interpolation est l'analogue de celle du polynôme d'interpolation sous forme de Lagrange. Comme dans le cas unidimensionnel on peut espérer qu'un choix heureux des points d'interpolation P_q et des fonctions $\omega_j(P)$ détermine une petite erreur dans les égalités approchées

$$f(P) \approx g(P) = \sum_{q=1}^{N} z_q(P) f(P_q), \tag{4}$$

$$\mathscr{D}f(P) \approx \mathscr{D}g(P) = \sum_{q=1}^{N} \mathscr{D}z_q(P) f(P_q), \tag{5}$$

$$I(f) \approx I(g) = \sum_{q=1}^{N} C_q f(P_q), \tag{6}$$

où

$$C_q = I(z_q),$$

$\mathscr{D}$ étant un opérateur de dérivation.

Comme nous l'avons observé (ch. II) pour le cas unidimensionnel, avec un mauvais choix de nombreux points d'interpolation l'erreur dans l'égalité approchée (4) peut s'avérer catastrophique. Les égalités (5), (6) découlant de l'égalité approchée (4), elles peuvent également être entachées d'erreurs importantes. C'est pourquoi l'application des relations (4), (5), (6) doit être justifiée et il faut éclaircir les conditions sous lesquelles leur emploi est licite.

§ 2. Méthode des moindres carrés

En construisant $g(P)$ on a essayé de la faire coïncider en N points avec la fonction $f(P)$. La fonction $g(P)$ contenant un grand nombre de paramètres, elle peut différer notablement de $f(P)$ en dehors de ces points.

Nous pouvons obtenir une meilleure approximation par des procédés différents. Considérons l'un d'eux, dit *méthode des moindres carrés*. On cherche l'approximation sous la forme

$$g(P) = \sum_{j=1}^{n} B_j \omega_j(P)$$

avec $n \ll N$; on définit les paramètres B_j par la condition

$$\min_{B_1, \ldots, B_n} \Phi(g)$$

avec, par exemple,

$$\Phi(g)=\Phi(B_1, \ldots, B_n)=\sum_{q=1}^{N} p_q (g(P_q)-f(P_q))^2=$$

$$=\sum_{q=1}^{N} p_q \Big(\sum_{j=1}^{n} B_j \omega_j(P_q)-f(P_q)\Big)^2. \qquad (1)$$

La méthode des moindres carrés est basée sur la considération ci-dessous. La proximité de $g(P)$ et $f(P)$ aux points P_q est assurée par la petitesse de $\Phi(g)$; si $n \ll N$, la fonction $g(P)$ contient un nombre relativement faible de paramètres, donc en dehors des points d'interpolation elle s'écarte moins de $f(P)$ que dans le cas $n = N$.

Le choix des nombres $p_q > 0$; qu'on appelle poids, dépend de la densité de répartition des points P_q ; si les valeurs $f(P_q)$ sont affectées d'une erreur aléatoire, il dépend également de la variance des erreurs sur les valeurs mesurées. Là où la densité de répartition des P_q est plus grande, on prend p_q plus petits; aux valeurs $f(P_q)$ avec une erreur de variance plus importante on fait correspondre des p_q plus faibles. Nos conseils paraissent assez vagues puisqu'on ne peut pas proposer une règle générale valable dans tous les cas. Pour des classes concrètes de problèmes on établit les principes régissant le choix de p_q et de n compte tenu des propriétés spécifiques des problèmes en partant des critères statistiques et de l'expérience numérique.

En égalant à zéro les dérivées $\frac{\partial \Phi}{\partial B_i}$ on obtient un système d'équations linéaires pour déterminer B_j:

$$\frac{1}{2}\frac{\partial \Phi}{\partial B_i}=\sum_{j=1}^{n} d_{ij}B_j-d_i=0,$$

$$d_{ij}=\sum_{q=1}^{N} p_q \omega_i(P_q)\,\omega_j(P_q), \qquad (2)$$

$$d_i=\sum_{q=1}^{N} p_q \omega_i(P_q)\,f(P_q).$$

On trouve les nombres B_j en résolvant directement ce système ou en minimisant d'une façon quelconque la fonction Φ. On transforme (1) en

$$\Phi(B_1, \ldots, B_n)=\sum_{i,\,j=1}^{n} d_{ij}B_iB_j-2\sum_{j=1}^{n} d_jB_j+d_0$$

avec

$$d_0=\sum_{q=1}^{N} p_q (f(P_q))^2.$$

Puisque $\Phi(B_1, \ldots, B_n) \geqslant 0$, la forme quadratique $\sum_{i,j=1}^{n} d_{ij}B_iB_j$ est non négative; on utilise cette propriété dans plusieurs programmes standard de la méthode des moindres carrés.

Exprimant B_j de (2) par l'intermédiaire de d_i, puis de $f(P_q)$, il vient

$$B_j = \sum_{q=1}^{N} a_{jq} f(P_q).$$

Donc

$$g(P) = \sum_{q=1}^{N} z_q(P) f(P_q) \tag{3}$$

avec

$$z_q(P) = \sum_{i=1}^{n} a_{iq}\omega_i(P).$$

En utilisant (3) on peut également obtenir la formule de dérivation numérique $\mathscr{D}f \approx \mathscr{D}g$ et la formule de quadrature

$$I(f) \approx I(g).$$

§ 3. Méthode de régularisation

La *méthode de régularisation* repose sur l'idée de lissage de la fonction approchante. Voici sa variante la plus répandue. L'approximation est cherchée sous la forme

$$g(P) = \sum_{j=1}^{n} B_j\omega_j(P),$$

on choisit les coefficients B_j tels que l'expression

$$\Phi(\lambda, g) = \Phi(g) + \lambda\Psi(g), \quad \lambda > 0, \tag{1}$$

soit minimale.

La fonctionnelle $\Psi(g)$ doit satisfaire à la condition suivante: si elle est petite, la fonction g est suffisamment régulière. Par exemple, $\Psi(g)$ peut être une approximation de $\int_G |\operatorname{grad} g(P)|^2 dP$. Nous nous bornons au cas $n = N$ dont les applications sont fréquentes. Supposons que le minimum de $\Phi(\lambda, g)$ est atteint pour certains $B_1^\lambda, \ldots, B_N^\lambda$ et

$$g^\lambda(P) = \sum_{j=1}^{N} B_j^\lambda\omega_j(P).$$

Considérons les cas extrêmes : $\lambda = 0$ et λ très grand. On a l'égalité

$$\Phi(0, g) = \sum_{j=1}^{N} p_q \left(\sum_{j=1}^{N} B_j \omega_j (P_q) - f(P_q) \right)^2.$$

Si $\det \| \omega_j (P_q) \| \neq 0$, le système (1.1) admet une solution qui annule le second membre de cette égalité. D'autre part, l'expression de $\Phi(0, g)$ est toujours non négative ; la borne inférieure est donc réalisée par des valeurs B_j qui sont solutions de (1.1). Alors $g^0(P)$ coïncide avec un polynôme d'interpolation basé sur les points P_j. Pour des λ grands dans $\Phi(\lambda, g)$ c'est le deuxième terme de la somme qui est prépondérant (sa borne inférieure est atteinte sur une fonction régulière). Nous pouvons donc espérer que pour les valeurs intermédiaires de λ les fonctions $g^\lambda(P)$ seront régulières tout en ne différant pas sensiblement de la fonction à approcher en des points donnés.

Nos raisonnements semblent manquer de rigueur, mais tel que le problème d'approximation est posé, sans que soient précisés le système de fonctions ω_j, la répartition des points P_j et la classe des fonctions considérées, on ne peut en escompter plus. De telles considérations « physiques » grossières sont souvent utiles lors de la construction de nouvelles méthodes de résolution d'un problème quand l'information sur celui-ci est insuffisante.

Un exemple concret du paragraphe suivant mettra en évidence l'effet de la régularisation.

§ 4. Exemple de régularisation

Soit $f(x)$ une fonction réelle périodique de période 1 ; on connaît des valeurs approchées f_q des $f(x_q)$ pour $x_q = qh$, $q = 0, \ldots, N-1$; $Nh = 1$. Les erreurs sur les valeurs approchées $f_q - f(x_q) = d_q$ sont des variables aléatoires d'espérance nulle et de variance d^2. On demande de dresser un tableau des valeurs approchées des dérivées $f'(x_q)$. On évalue l'erreur par une norme associée au produit scalaire

$$(f, g) = \left(\sum_{q=0}^{N-1} f_q \bar{g}_q \right) h.$$

f_q est considérée ensuite comme périodiquement prolongée.

On trouve la dérivée en utilisant la formule simple de dérivation numérique

$$f'(x_q) \approx \mathcal{D} f_q = \frac{f_{q+1} - f_{q-1}}{2h}. \tag{1}$$

Supposons que la fonction $f(x)$ soit suffisamment régulière et le pas h si petit que, en l'absence d'erreurs sur les valeurs de la fonction, l'erreur dans la formule de dérivation numérique

$$f'(x_q) \approx \frac{f(x_{q+1}) - f(x_{q-1})}{2h} \tag{2}$$

est négligeable. L'erreur sur la valeur de la dérivée se met sous la forme suivante :

$$R_q = \frac{f_{q+1} - f_{q-1}}{2h} - f'(x_q) =$$

$$= \left(\frac{f(x_{q+1}) - f(x_{q-1})}{2h} - f'(x_q) \right) + \frac{d_{q+1} - d_{q-1}}{2h} \approx \frac{d_{q+1} - d_{q-1}}{2h} = r_q.$$

Ici et plus loin le signe d'égalité approchée signifie l'égalité à l'ordre de grandeur de l'erreur sur la formule (2) près. On a

$$\mathrm{E} \| R \|^2 \approx \mathrm{E} \| r \|^2 = \mathrm{E} \left(\sum_{q=0}^{N-1} \frac{d_{q+1} - d_{q-1}}{2h} \right)^2 h \Big) =$$

$$= \frac{N}{4} \sum_{q=0}^{N-1} (\mathrm{E} d_{q+1}^2 - 2\mathrm{E}(d_{q-1} d_{q+1}) + \mathrm{E} d_{q-1}^2).$$

Conformément aux propriétés indiquées des variables aléatoires d_q,

$$\mathrm{E} d_q d_p = d^2 \delta_{p-q}. \qquad (3)$$

On obtient alors $\mathrm{E} \| R \|^2 = N^2 d^2/2$, autrement dit, la moyenne quadratique de la norme de l'erreur est égale à $Nd/\sqrt{2}$. On voit que l'erreur dans la formule approchée (1) s'accroît avec la diminution de $h = 1/N$; d'autre part, la petitesse de l'erreur sous l'hypothèse de l'absence d'erreurs sur (2) exige un h suffisamment petit.

Voyons comment réduire l'erreur. Soient A_j^0 les coefficients de Fourier discrets (c.F.d.) de la fonction $f(x_q)$:

$$f(x_q) = \sum_{-N/2 < j \leqslant N/2} A_j^0 \exp(2\pi i j x_q),$$

A_j étant un c.F.d. de f_q. Les quantités $a_j = A_j - A_j^0$ seront les c.F.d. de la fonction $d_q = f_q - f(x_q)$. On a

$$d_q = \sum_{-N/2 < j \leqslant N/2} a_j \exp(2\pi i j x_q), \qquad (4)$$

$$a_j = h \sum_{q=0}^{N-1} d_q \exp(-2\pi i j x_q).$$

Pour calculer les c.F.d. de r_q, on applique l'opérateur $\mathscr{D}$ à la fonction $\exp(2\pi i j x_q)$:

$$\mathscr{D} \exp(2\pi i j x_q) = \frac{\exp(2\pi i j x_{q+1}) - \exp(2\pi i j x_{q-1})}{2h} = \frac{i \sin 2\pi j h}{h} \exp(2\pi i j x_q).$$

En appliquant l'opérateur $\mathscr{D}$ aux deux membres de (4), il vient

$$r_q = \sum_{-N/2 < j \leqslant N/2} \frac{i \sin 2\pi j h}{h} a_j \exp(2\pi i j x_q).$$

Recourons à l'égalité de Parseval :

$$\| r \|^2 = \sum_{-N/2<j\leqslant N/2} \left(\frac{\sin 2\pi jh}{h}\right)^2 |a_j|^2.$$

On en déduit :

$$\mathrm{E}\,\| r \|^2 = \sum_{-N/2<j\leqslant N/2} \left(\frac{\sin 2\pi jh}{h}\right)^2 \mathrm{E}\,|a_j|^2.$$

On a

$$|a_j|^2 = a_j\bar{a}_j = h^2 \sum_{q,\,p=0}^{N-1} d_q d_p \exp\left(2\pi ij\,(x_p - x_q)\right);$$

nous avons remplacé $\bar{d}_p$ par d_p ayant utilisé le fait que les erreurs sur d_p sont réelles. Alors,

$$\mathrm{E}\,|a_j|^2 = h^2 \sum_{q,\,p=0}^{N-1} \exp\left(2\pi ij\,(x_p - x_q)\right) \mathrm{E}\,(d_q d_p) = h^2 \sum_{q,\,p=0}^{N-1} d^2\delta_{p-q} = hd^2.$$

Donc,

$$\mathrm{E}\,\| r \|^2 = \sum_{-N/2<j\leqslant N/2} \left(\frac{\sin 2\pi jh}{h}\right)^2 hd^2. \tag{5}$$

Prenons une fonction $\mu(j)$ et posons

$$f^{\mu}(x) = \sum_{-N/2<j\leqslant N/2} A_j\mu(j) \exp(2\pi ijx),$$

$$f_q^{\mu} = \sum_{-N/2<j\leqslant N/2} A_j\mu(j) \exp(2\pi ijx_q).$$

Pour des valeurs f_q données on peut calculer les coefficients A_j, donc f_q^{μ}. Soit r_q^{μ} un terme de l'erreur dans la formule de dérivation numérique qui est le résultat des erreurs sur d_q :

$$r_q^{\mu} = \mathscr{D} f_q^{\mu} - \mathscr{D} f^{\mu}(x_q).$$

D'une façon analogue à (5) on obtient

$$\mathrm{E}\,\| r^{\mu} \|^2 = \sum_{-N/2<j\leqslant N/2} \left(\frac{\sin 2\pi jh}{h}\right)^2 \mu^2(j)\, hd^2;$$

si

$$\mathscr{D} f(x_q) \approx \mathscr{D} f^{\mu}(x_q), \quad \mathrm{E}\,\| r^{\mu} \|^2 \ll \mathrm{E}\,\| r \|^2, \tag{6}$$

il est raisonnable de poser $f'(x_q) \approx \mathscr{D} f_q^{\mu}$. La validité de la première relation (6) dépend essentiellement de la proximité des coefficients de Fourier de $f(x)$ et de $f^{\mu}(x)$, celle de la deuxième de la petitesse de $\mu(j)$ pour les j grands.

Tâchons de lisser la fonction f_q par la méthode de régularisation. Considérons la fonctionnelle

$$\Phi_1^h(\lambda, g) = h \sum_{q=0}^{N-1} (g_q - f_q)^2 + \lambda^2 h \sum_{q=0}^{N-1} \left(\frac{g_{q+1} - g_q}{h}\right)^2$$

avec $g_N = g_0$. Pour $g_q = g(x_q)$, $g(x_q)$ régulière, la quantité

$$h = \sum_{q=0}^{N-1} \left(\frac{g_{q+1} - g_q}{h} \right)^2$$

tend vers l'intégrale $\int_0^1 (g'(x))^2 \, dx$; si cette intégrale n'est pas grande, la fonction $g(x)$ est suffisamment régulière. Il y a donc des raisons de croire que la fonctionnelle $\Phi_1^h(\lambda, g)$ satisfait aux conditions imposées aux fonctionnelles de la méthode de régularisation. Déterminons une fonction discrète g_q^λ à partir de la condition de minimum de la fonctionnelle $\Phi_1^h(\lambda, g)$ et posons

$$f'(x_q) \approx \frac{g_{q+1}^\lambda - g_{q-1}^\lambda}{2h}.$$

Voyons quel est l'analogue de cette régularisation dans le cas de fonctions d'une variable continue. Soit

$$f(x) = \sum_{j=-\infty}^{\infty} A_j^0 \exp(2\pi i j x)$$

et $g_1^\lambda(x)$ une fonction réalisant le minimum de la fonctionnelle

$$\Phi_1(\lambda, g) = \int_0^1 (g(x) - f(x))^2 \, dx + \lambda^2 \int_0^1 (g'(x))^2 \, dx.$$

L'équation d'Euler pour cette fonctionnelle est

$$\lambda^2 g'' - (g - f) = 0.$$

Un calcul direct montre que la fonction

$$g_1^\lambda(x) = \sum_{j=-\infty}^{\infty} \frac{A_j^0}{1 + (\lambda 2\pi j)^2} \exp(2\pi i j x)$$

est solution de cette équation. Comparons $g_1^\lambda(x)$ et $f(x)$ pour les λ petits. Si $\lambda 2\pi j \ll 1$, les coefficients de Fourier $\frac{A_j^0}{1 + (\lambda 2\pi j)^2}$ et A_j^0 de ces fonctions sont proches. Si $\lambda 2\pi j \gg 1$, les coefficients de Fourier de la fonction $g_1^\lambda(x)$ sont sensiblement inférieurs à ceux de $f(x)$. Donc la régularisation équivaudrait en termes de technique à une « filtration »: en déformant légèrement les harmoniques d'oscillation à basse fréquence, elle atténue considérablement ceux à haute fréquence. S'il faut déformer encore moins les amplitudes A_j^0 des harmoniques à basse fréquence et « filtrer » plus fortement les oscil-

lations à haute fréquence, on considérera la fonctionnelle

$$\Phi_n(\lambda, g) = \int_0^1 (g(x) - f(x))^2 dx + \lambda^{2n} \int_0^1 (g^{(n)}(x))^2 dx.$$

L'équation d'Euler pour cette fonctionnelle est

$$\lambda^{2n} g^{(2n)} + (-1)^n (g - f) = 0,$$

d'où

$$g_n^\lambda(x) = \sum_{j=-\infty}^{\infty} \frac{A_j^0}{1 + (\lambda 2\pi j)^{2n}} \exp(2\pi i j x).$$

Considérons les courbes des facteurs $\frac{1}{1 + (\lambda 2\pi j)^{2n}}$; pour $|j| < \frac{1}{2\pi\lambda}$ et $n \to \infty$ ces derniers tendent vers 1, et, donc, pour les grands n

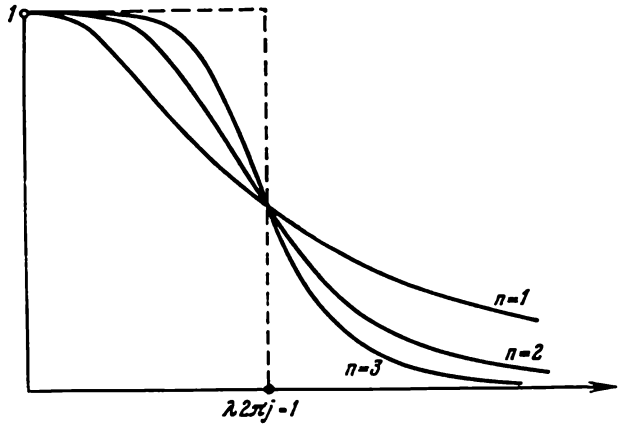


Fig. 5.4.1

les amplitudes des harmoniques correspondants sont déformées de moins en moins. D'autre part, pour $|j| > \frac{1}{2\pi\lambda}$, ces facteurs tendent vers 0, et les harmoniques correspondants sont multipliés par des quantités toujours plus petites (fig. 5.4.1). On a donc pour $n \to \infty$ et $\frac{1}{2\pi\lambda}$ non entier

$$g_n^\lambda(x) \to \tilde{g}^\lambda(x) = \sum_{|j| < 1/(2\pi\lambda)} A_j^0 \exp(2\pi i j x).$$

Reprenons le cas discret. Dans l'expression de $\Phi_1^h(\lambda, g)$ la quantité g_q ne figure que dans les termes

$$h(g_q - f_q)^2 + \lambda^2 h \left(\left(\frac{g_{q+1} - g_q}{h} \right)^2 + \left(\frac{g_q - g_{q-1}}{h} \right)^2 \right).$$

On a le système d'équations par rapport aux valeurs g_q au point d'extrémum

$$\frac{1}{2h}\frac{\partial \Phi_1^h}{\partial g_q} = (g_q - f_q) - \frac{\lambda^2}{h^2}(g_{q+1} - 2g_q + g_{q-1}) = 0. \tag{7}$$

La matrice du système (7) relativement aux inconnues g_q coïncide avec celle de la forme quadratique définie positive

$$\sum_{q=0}^{N-1} g_q^2 + \frac{\lambda^2}{h^2}\sum_{q=0}^{N-1}(g_{q+1} - g_q)^2;$$

donc, son déterminant est non nul et le système (7) a une solution et une seule.

Soit

$$f_q = \sum_{-N/2<j\leqslant N/2} A_j \exp(2\pi i j x_q);$$

cherchons la fonction périodique g_q sous la forme

$$g_q^\lambda = \sum_{-N/2<j\leqslant N/2} A_j^\lambda \exp(2\pi i j x_q).$$

Calculons d'abord l'opérateur différence deuxième de la fonction $\exp(2\pi i j x_q)$:

$$\begin{aligned}\delta^2 \exp(2\pi i j x_q) &= \exp(2\pi i j x_{q+1}) - 2\exp(2\pi i j x_q) + \exp(2\pi i j x_{q-1}) = \\ &= (\exp(2\pi i j h) - 2 + \exp(-2\pi i j h))\exp(2\pi i j x_q) = \\ &= (2\cos(2\pi j h) - 2)\exp(2\pi i j x_q),\end{aligned}$$

d'où

$$\delta^2 \exp(2\pi i j x_q) = -4\sin^2(\pi j h)\exp(2\pi i j x_q). \tag{8}$$

Substituons dans (7) la représentation de f_q et g_q sous la forme de somme de Fourier, transformons la différence deuxième $\delta^2 g_q$ compte tenu de (8) et réduisons les termes semblables :

$$\sum_{-N/2<j\leqslant N/2}\left(A_j^\lambda - A_j + 4\lambda^2\left(\frac{\sin \pi j h}{h}\right)^2 A_j^\lambda\right)\exp(2\pi i j x_q) = 0.$$

Cette égalité est satisfaite pour

$$A_j^\lambda = \frac{A_j}{1 + 4\lambda^2\left(\frac{\sin \pi j h}{h}\right)^2}.$$

Donc,

$$g_q^\lambda = \sum_{-N/2<j\leqslant N/2} \frac{A_j}{1 + 4\lambda^2\left(\frac{\sin \pi j h}{h}\right)^2}\exp(2\pi i j x_q).$$

Le facteur

$$\frac{1}{1+4\lambda^2\left(\frac{\sin \pi jh}{h}\right)^2}$$

décroît lorsque j croît. On voit que la régularisation au moyen de la fonctionnelle $\Phi_1^h(\lambda, g)$ conduit également à l'atténuation des harmoniques supérieurs de la fonction. Par analogie avec les fonctionnelles $\Phi_n(\lambda, g)$ on peut effectuer une régularisation à l'aide des fonctionnelles

$$\Phi_n^h(\lambda, g) = h\sum_{q=0}^{N-1}(g_q - f_q)^2 + \lambda^{2n}h\sum_{q=0}^{N-1}\left(\frac{\Delta^n g_q}{h^n}\right)^2.$$

On résout le système (7) et les systèmes d'équations linéaires qui apparaissent en minimisant les fonctionnelles $\Phi_n^h(\lambda, g)$ au bout de $O(n^2N)$ opérations arithmétiques en appliquant la méthode du balayage au problème aux limites périodique discrétisé (voir § 3 ch. IX). Pour $n > 1$ on préférera une régularisation au moyen des fonctionnelles $\Phi_n^h(\lambda, g)$ à une régularisation n-uple par la fonctionnelle $\Phi_1^h(\lambda, g)$: dans ce cas les harmoniques $\exp(2\pi ijx)$ sont multipliés par

$$\frac{1}{\left(1+4\lambda^2\left(\frac{\sin \pi jh}{h}\right)^2\right)^n}.$$

L'avantage d'une telle régularisation successive est la simplicité de l'algorithme, un volume de calculs moindre (de l'ordre de $O(nN)$) et une influence moins importante de l'erreur de calcul.

Dans le cas considéré, on peut procéder autrement: on effectue une transformation de Fourier discrète (rapide si possible), on trouve les coefficients de Fourier A_j, on substitue 0 aux A_j correspondant aux harmoniques haute fréquence, puis on effectue une transformation de Fourier inverse.

§ 5. Réduction des problèmes multidimensionnels aux problèmes unidimensionnels

Les procédés de résolution des problèmes multidimensionnels étudiés jusqu'ici n'utilisaient pas l'information supplémentaire sur la répartition des points où les valeurs de la fonction sont connues.

Dans de nombreux cas la disposition de ces points est régulière; si la fonction en question est donnée analytiquement, on les choisit de façon arbitraire. En cas d'une disposition heureuse on ramène les opérations d'approximation, d'interpolation, de dérivation et d'intégration numériques des fonctions de plusieurs variables à celles effectuées successivement sur des fonctions d'une variable.

La plupart de ces cas obéissent au schéma suivant. Supposons connues les valeurs d'une fonction sur un ensemble de points. On peut faire un changement de variables par lequel les valeurs de la fonction $f(y_1, \ldots, y_s)$ seront données sur l'ensemble Ω des points ayant les propriétés suivantes: l'ensemble Ω se décompose en n_1 ensembles $\Omega_1, \ldots, \Omega_{n_1}$ de façon que la plupart des ensembles Ω_{m_1} contiennent des points en nombre suffisamment grand; Ω_{m_1} se trouve dans l'hyperplan $y_1 = y_1^{m_1}$, $m_1 = 1, \ldots, n_1$; à son tour, l'ensemble Ω_{m_1} se décompose en ensembles $\Omega_{m_1 m_2}$; $m_2 = 1, \ldots$ $\ldots, n_2(m_1)$, tels que la plupart d'entre eux contiennent un nombre

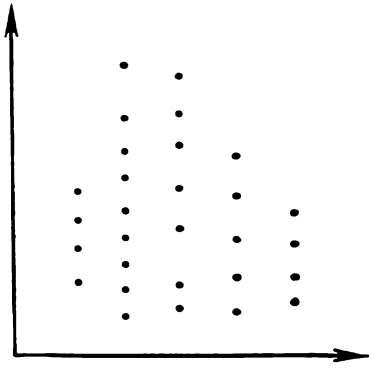


Fig. 5.5.1

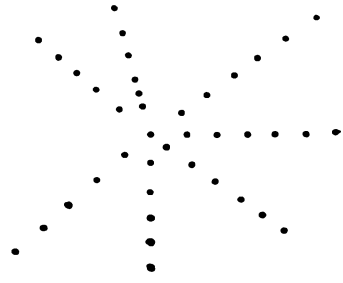


Fig. 5.5.2

suffisamment grand de points, l'ensemble $\Omega_{m_1 m_2}$ se trouve dans le plan $y_1 = y_1^{m_1}$, $y_2 = y_2^{m_1 m_2}$, $\Omega_{m_1 m_2}$ se décompose à son tour en ensembles $\Omega_{m_1 m_2 m_3}$, et ainsi de suite. Autrement dit, tous les points peuvent s'écrire sous la forme

$$(y_1^{m_1}, y_2^{m_1 m_2}, \ldots, y_s^{m_1 \ldots m_s}), \tag{1}$$

où

$$1 \leqslant m_1 \leqslant n_1,\ 1 \leqslant m_k \leqslant n_k(m_1, \ldots, m_{k-1}) \text{ pour } k > 1.$$

Il est clair que tout ensemble de points peut être mis sous la forme (1); cependant nous exigeons de plus que les ensembles $\Omega_{m_1 \ldots m_k}$ comprennent des points en nombre suffisamment grand. Des exemples de tels ensembles sont représentés fig. 5.5.1, 5.5.2. Le cas correspondant à la fig. 5.5.2. satisfait aux conditions ci-dessus relatives au système de coordonnées polaires. Soit à calculer la valeur d'un opérateur

$$\mathcal{D}f(P^0) = f^{(r_1, \ldots, r_s)}(y_1^0, \ldots, y_s^0);$$

le cas particulier $r_1 = \ldots = r_s = 0$ correspond au problème d'interpolation. Prenons une formule quelconque de calcul de la dérivée $g^{(r_1)}(y_1^0)$ d'après les valeurs de la fonction aux points $y_1^{m_1}$:

$$g^{(r_1)}(y_1^0) \approx \sum_{m_1=1}^{n_1} D^{m_1} g(y_1^{m_1}).$$

Substituant

$$g(y_1) = f^{(0, r_2, \ldots r_s)}(y_1, y_2^0, \ldots y_s^0)$$

on en tire

$$f^{(r_1, \ldots, r_s)}(y_1^0, \ldots, y_s^0) \approx \sum_{m_1=1}^{n_1} D^{m_1} f^{(0, r_2, \ldots, r_s)}(y_1^{m_1}, y_2^0, \ldots, y_s^0). \quad (2)$$

La dérivation de la fonction $f(y_1, \ldots, y_s)$ de s variables au moyen de cette égalité se ramène à dériver les fonctions $f(y_1^{m_1}, y_2, \ldots, y_s)$ de $s-1$ variables. On connaît les valeurs de ces fonctions sur les réseaux Ω_{m_1} satisfaisant aux mêmes conditions que le réseau initial.

Pour chaque m_1 choisissons une formule de calcul de la dérivée $g^{(r_2)}(y_2^0)$ à partir des valeurs de la fonction aux points $y_2^{m_1 m_2}$:

$$g^{(r_2)}(y_2^0) \approx \sum_{m_2=1}^{n_2(m_1)} D^{m_1 m_2} g(y_2^{m_1 m_2}).$$

En y substituant

$$g = f^{(0, 0, r_3, \ldots, r_s)}(y_1^{m_1}, y_2, y_3^0, \ldots, y_s^0),$$

on obtient, d'une façon analogue à (2),

$$f^{(0, r_2, \ldots, r_s)}(y_1^{m_1}, y_2^0, \ldots, y_s^0) \approx$$

$$\approx \sum_{m_2=1}^{n_2(m_1)} D^{m_1 m_2} f^{(0, 0, r_3, \ldots, r_s)}(y_1^{m_1}, y_2^{m_1 m_2}, y_3^0, \ldots, y_s^0).$$

En utilisant cette relation il vient de (2)

$$f^{(r_1, \ldots, r_s)}(y_1^0, \ldots, y_s^0) \approx$$

$$\approx \sum_{m_1=1}^{n_1} D^{m_1} \Big(\sum_{m_2=1}^{n_2(m_1)} D^{m_1 m_2} f^{(0, 0, r_3, \ldots, r_s)}(y_1^{m_1}, y_2^{m_1 m_2}, y_3^0, \ldots, y_s^0) \Big),$$

etc. En poursuivant cette procédure, on obtient définitivement la formule de dérivation numérique

$$f^{(r_1, \ldots, r_s)}(y_1^0, \ldots, y_s^0) \approx \sum_{m_1=1}^{n_1} D^{m_1} \sum_{m_2=1}^{n_2(m_1)} D^{m_1 m_2} \ldots$$

$$\ldots \sum_{m_s=1}^{n_s(m_1, \ldots, m_{s-1})} D^{(m_1, \ldots, m_s)} f^{(0, \ldots, 0)}(y_1^{m_1}, \ldots, y_s^{m_1, \ldots, m_s}), \quad (3)$$

où $f^{(0, \ldots, 0)} \equiv f$. Le fait que le deuxième membre comprenne la somme des termes en tous les points ne doit pas inquiéter: en réalité, les formules unidimensionnelles de dérivation numérique et d'interpolation qui nous ont permis d'obtenir (3) contiennent relativement peu de points.

De la même façon on construit les formules d'intégration numérique.

Soit à calculer l'intégrale

$$I_0 = \int_G f(y_1, \ldots, y_s)\, dy_1, \ldots, dy_s.$$

On peut la mettre sous la forme

$$I_0 = \int_{G_0} I_1(y_1)\, dy_1, \quad I_1(y_1) = \int_{G_1(y_1)} I_2(y_1, y_2)\, dy_2,$$

$$I_2(y_1, y_2) = \int_{G_2(y_1, y_2)} I_3(y_1, y_2, y_3)\, dy_3, \ldots, I_{s-1}(y_1, \ldots, y_{s-1}) =$$

$$= \int_{G_{s-1}(y_1, \ldots, y_{s-1})} I_s(y_1, \ldots, y_s)\, dy_s,$$

où

$$I_s(y_1, \ldots, y_s) = f(y_1, \ldots, y_s),$$

G_0 est la projection du domaine G sur l'axe y_1, $G_k(\bar{y}_1, \ldots, \bar{y}_k)$ la section de l'ensemble G par le plan $y_1 = \bar{y}_1, \ldots, y_k = \bar{y}_k$. Nous utilisons une formule de quadrature quelconque pour calculer la première de ces intégrales, à savoir

$$I_0 \approx \sum_{m_1=1}^{n_1} D^{m_1} I_1(y_1^{m_1}).$$

Le calcul de l'intégrale initiale s'est réduit au calcul d'intégrales de dimension moins 1. Posons maintenant

$$I_1(y_1^{m_1}) \approx \sum_{m_2=1}^{n_2(m_1)} D^{m_1 m_2} I_2(y_1^{m_1}, y_2^{m_1 m_2}),$$

$$\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots$$

$$I_{s-1}(y_1^{m_1}, \ldots, y_{s-1}^{m_1, \ldots, m_{s-1}}) \approx$$

$$\approx \sum_{m_s=1}^{n_s(m_1, \ldots, m_{s-1})} D^{m_1, \ldots, m_s} f(y_1^{m_1}, \ldots, y_s^{m_1, \ldots, m_s}). \tag{4}$$

On a définitivement

$$I_0 \approx \sum_{m_1=1}^{n_1} D^{m_1} \cdot \sum_{m_2=1}^{n_2(m_1)} D^{m_1 m_2} \ldots$$

$$\ldots \sum_{m_s=1}^{n_s(m_1, \ldots, m_{s-1})} D^{m_1, \ldots, m_s} f(y_1^{m_1}, \ldots, y_s^{m_1, \ldots, m_s}). \qquad (5)$$

Dans le cas unidimensionnel, pour toutes les opérations de l'analyse numérique on a évalué l'erreur au moyen des dérivées de la fonction considérée. Voyons les estimations qui en découlent pour les problèmes multidimensionnels. La valeur d'un opérateur de la fonction est approchée par des valeurs d'autres opérateurs, l'erreur sur un tel changement pouvant être évaluée selon les relations appréciant l'erreur dans les formules unidimensionnelles. Or, ces nouvelles valeurs contiennent elles-mêmes des erreurs ; c'est pourquoi celles-ci figureront dans l'erreur totale multipliées par certains facteurs. Reprenons notre problème d'intégration ; soit

$$I_{k-1}(y_1^{m_1}, \ldots, y_{k-1}^{m_1, \ldots, m_{k-1}}) =$$

$$= \sum_{m_k=1}^{n_k(m_1, \ldots, m_{k-1})} D^{m_1, \ldots, m_k} I_k(y_1^{m_1}, \ldots, y_k^{m_1, \ldots, m_k}) + E_{m_1, \ldots, m_{k-1}}^{k-1}.$$

En substituant successivement les expressions de $I_1, I_2, \ldots$ dans l'égalité

$$I_0 = \sum_{m_1=1}^{n_1} D^{m_1} I_1(y_1^{m_1}) + E^0$$

on obtient la chaîne de relations

$$I_0 = E^0 + \sum_{m_1=1}^{n_1} D^{m_1} I_1(y_1^{m_1}) =$$

$$= E^0 + \sum_{m_1=1}^{n_1} D^{m_1} \left(E_{m_1}^1 + \sum_{m_2=1}^{n_2(m_1)} D^{m_1 m_2} I_2(y_1^{m_1}, y_2^{m_1 m_2})\right) = \ldots$$

$$\ldots = R + \sum_{m_1=1}^{n_1} D^{m_1} \Big(\sum_{m_2=1}^{n_2(m_1)} D^{m_1 m_2} \times$$

$$\times \Big(\ldots \Big(\sum_{m_s=1}^{n_s(m_1, \ldots, m_{s-1})} D^{m_1, \ldots, m_s} f(y_1^{m_1}, \ldots, y_s^{m_1, \ldots, m_s}) \Big) \ldots \Big) \Big)$$

avec

$$R = E^0 + \sum_{m_1=1}^{n_1} D^{m_1} E^1_{m_1} + \sum_{m_1=1}^{n_1} D^{m_1} \sum_{m_2=1}^{n_2(m_1)} D^{m_1 m_2} E^2_{m_1 m_2} + \ldots$$
$$\ldots + \sum_{m_1=1}^{n_1} D^{m_1} \sum_{m_2=1}^{n_2(m_1)} D^{m_1 m_2} \ldots$$
$$\ldots \sum_{m_{s-1}=1}^{n_{s-1}(m_1, \ldots, m_{s-2})} D^{m_1, \ldots, m_{s-1}} E^{s-1}_{m_1, \ldots, m_{s-1}}. \qquad (6)$$

La dernière égalité montre que si les coefficients $D^{m_1, \ldots, m_k}$ sont grands, l'erreur d'approximation peut devenir considérablement plus importante que dans le cas unidimensionnel.

Formellement, la réduction d'un problème multidimensionnel à un autre de dimension 1 est identique qu'il s'agisse de la dérivation numérique (interpolation) ou de l'intégration numérique. On note cependant une différence essentielle: la dérivation numérique (interpolation) consiste le plus souvent à chercher l'opérateur d'une fonction à partir des valeurs prises sur un ensemble de points Ω donné. Tandis que le problème d'intégration se caractérise par la possibilité de choisir à loisir les points.

Considérons un exemple de constructions pareilles. Nous voudrions attirer l'attention du lecteur sur la nécessité d'être prudent en choisissant les formules unidimensionnelles auxiliaires d'intégration numérique et souligner l'importance pratique de l'évaluation du reste. Supposons que les valeurs d'une fonction $f(x_1, x_2)$ sont connues aux points $(m_1 h, m_2 h)$ du réseau. Soit à calculer la valeur $f_{x_1 x_2}(0, 0)$. Ecrivons une formule simple de dérivation numérique de la fonction d'une variable:

$$f'(x) \sim \frac{f(x+h) - f(x)}{h}.$$

Vu la formule de Taylor

$$f(x+h) = f(x) + f'(x)\, h + f''(x+\theta h)\frac{h^2}{2}, \quad 0 \leqslant \theta \leqslant 1,$$

on a

$$f'(x) = \frac{f(x+h) - f(x)}{h} - f''(x+\theta h)\frac{h}{2}.$$

Donc, nous pouvons écrire

$$f_{x_1 x_2}(0, 0) = \frac{f_{x_2}(h, 0) - f_{x_2}(0, 0)}{h} - f_{x_1^2 x_2}(\theta h, 0)\frac{h}{2}. \qquad (7)$$

Prenons, d'autre part, des approximations des dérivées $f_{x_2}(h, 0)$, $f_{x_2}(0, 0)$; on a les égalités

$$f_{x_2}(h, 0) = \frac{f(h, h) - f(h, 0)}{h} - f_{x_2x_2}(h, \theta_1 h)\frac{h}{2},$$

$$f_{x_2}(0, 0) = \frac{f(0, 0) - f(0, -h)}{h} + f_{x_2x_2}(0, \theta_2 h)\frac{h}{2}.$$

Après substitution dans (7), on obtient

$$f_{x_1x_2}(0, 0) = \frac{f(h, h) - f(h, 0) - f(0, 0) + f(0, -h)}{h^2} -$$

$$- f_{x_1^2x_2}(\theta h, 0)\frac{h}{2} - \frac{1}{2} f_{x_2x_2}(0, \theta_1 h) - \frac{1}{2} f_{x_2x_2}(h, \theta_2 h). \quad (8)$$

En construisant l'approximation sans tenir compte du reste on aurait été amené à la conclusion fausse que l'expression

$$\frac{f(h, h) - f(h, 0) - f(0, 0) + f(0, -h)}{h^2}$$

est une bonne approximation de la valeur $f_{x_1x_2}(0, 0)$. On peut récrire la relation (8) comme suit:

$$f_{x_1x_2}(0, 0) = \frac{f(h, h) - f(h, 0) - f(0, 0) + f(0, -h)}{h^2} - f_{x_2x_2}(0, 0) + O(h).$$

Ainsi, si on néglige la croissance éventuelle de l'erreur due à la multiplication par les $D^{m_1 \cdots m_k}$ grands, on peut obtenir une approximation affectée d'une erreur finie, qui approche dans notre cas non

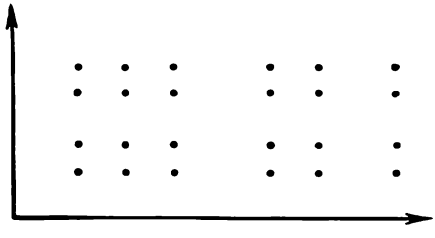


Fig. 5.5.3

pas la dérivée cherchée mais l'expression $f_{x_1x_2}(0, 0) + f_{x_2x_2}(0, 0)$.

Le cas où les points forment un réseau proprement dit, en d'autres termes, l'ensemble des points de la forme

$$(y_1^{m_1}, \ldots, y_s^{m_s}) \text{ pour } m_1 = 1, \ldots, N_1; \ldots; m_s = 1, \ldots, N_s$$

est important dans les applications. La figure 5.5.3 représente une telle disposition des points pour $s = 2$. Dans ce cas, pour réaliser

les opérations multidimensionnelles de dérivation numérique (d'interpolation et d'intégration numérique), il est raisonnable d'utiliser les mêmes formules pour approximer les valeurs intermédiaires. Cela signifie que les formules (4) par exemple doivent s'écrire:

$$I_{k-1}(y_1^{m_1}, \ldots, y_{k-1}^{m_{k-1}}) \approx \sum_{m_k=1}^{N_k} D^{m_k} I_k (y_1^{m_1}, \ldots, y_k^{m_k}), \tag{9}$$

c'est-à-dire que $D^{m_1 \cdots m_k}$ ne dépendent que de m_k. Le second membre de (5) se met alors sous la forme

$$\sum_{m_1=1}^{N_1} \cdots \sum_{m_s=1}^{N_s} D^{m_1} \ldots D^{m_s} f(y_1^{m_1}, \ldots, y_s^{m_s}).$$

En construisant une telle formule de quadrature nous supposons implicitement que le domaine d'intégration est un parallélépipède rectangle d'arêtes parallèles aux axes de coordonnées.

Ces formules d'intégration numérique (d'interpolation, de dérivation numérique) sont dites *produit direct* des formules unidimensionnelles correspondantes d'intégration numérique (d'interpolation, de dérivation numérique).

Un exemple plus compliqué du produit direct de la formule de quadrature sur un segment par la formule sur une sphère sera donné au § 13. En utilisant ces approximations on constate que les restes dans le second membre de (6) sont largement compensés. Dans le même exemple, en utilisant les formules du type (9) on obtient l'approximation

$$f_{x_1x_2}(0, 0) \approx \frac{f(h, h) - f(0, h) - f(h, 0) + f(0, 0)}{h^2}$$

avec une erreur de l'ordre de $O(h)$.

Il peut arriver dans l'intégration que pour une fonction régulière sous le signe somme les intégrales intermédiaires $I_k(y_1, \ldots, y_k)$ ne le soient pas suffisamment.

Soit à calculer l'intégrale sur le cercle unité

$$I = \int_{-1}^{1} \left(\int_{-\sqrt{1-y_1^2}}^{\sqrt{1-y_1^2}} f(y_1, y_2)\, dy_2 \right) dy_1$$

pour $f(y_1, y_2)$ régulière ; si $f(\pm 1, 0) \neq 0$, la fonction

$$I_1(y_1) = \int_{-\sqrt{1-y_1^2}}^{\sqrt{1-y_1^2}} f(y_1, y_2)\, dy_2$$

a ses dérivées non bornées aux points ± 1 ; donc, en intégrant numériquement en y_1 il faut utiliser des méthodes spéciales pour calculer les intégrales de telles fonctions: un pas variable, l'isolement des singularités, etc.; il vaut mieux écrire cette intégrale sous la forme

$$I = \int_0^1 r I_1(r)\, dr, \quad I_1(r) = \int_0^{2\pi} f(r\cos\varphi, r\sin\varphi)\, d\varphi.$$

Comme la fonction sous le signe somme intérieur est périodique, on a intérêt à utiliser la formule de quadrature (3.3.16)

$$\int_0^{2\pi} g(\varphi)\, d\varphi \approx \frac{2\pi}{n} \sum_{m=0}^{n-1} g\left(\frac{2\pi m}{n}\right).$$

Pour r faible, les dérivées de f par rapport à φ contiennent un petit facteur r; aussi pour φ peu élevés il faut parfois prendre n plus petit. Pour calculer l'intégrale en r on utilise n'importe quelles formules de quadrature examinées plus haut et notamment celles de Gauss correspondant à la fonction de poids r.

§ 6. Evaluation de l'erreur dans l'intégration numérique sur un réseau régulier

A titre d'exemple de procédés concrets d'évaluation des erreurs résultant des méthodes décrites ci-dessus, considérons le calcul de l'intégrale sur le cube unité:

$$I = \int_0^1 \dots \int_0^1 f(x_1, \dots, x_s)\, dx_1 \dots dx_s.$$

Supposons que la fonction sous le signe somme appartienne à la classe

$$C_r(\mathrm{A}) = C_{r_1 \dots r_s}(A_1, \dots, A_s)$$

des fonctions $f(x_1, \dots, x_s)$ dont les dérivées $f_{x_k^{r_k-1}}(x_1, \dots, x_s)$ sont continues et les dérivées $f_{x_k^{r_k}}(x_1, \dots, x_s)$ continues par morceaux et satisfont aux conditions $|f_{x_k^{r_k}}(x_1, \dots, x_s)| \leqslant A_k$. On considérera des formules de quadrature produit direct des formules unidimensionnelles. Posons

$$I_s(x_1, \dots, x_s) = f(x_1, \dots, x_s)$$

et

$$I_k(x_1, \dots, x_k) = \int_0^1 \dots \int_0^1 f(x_1, \dots, x_s)\, dx_{k+1} \dots dx_s$$

pour $k = 0, \dots, s-1$.

On a les inégalités

$$|(I_k)_{x_k^{r_k}}| \leqslant \int_0^1 \cdots \int_0^1 |f_{x_k^{r_k}}(x_1, \ldots, x_s)|\, dx_{k+1} \ldots dx_s \leqslant A_k.$$

Pour calculer les intégrales

$$I_{k-1}(x_1, \ldots, x_{k-1}) = \int_0^1 I_k(x_1, \ldots, x_k)\, dx_k$$

on utilisera une formule composite (3.6.4) avec $n = r_k$ et un pas constant $a_q - a_{q-1} \equiv H_k = N_k^{-1}$:

$$\int_0^1 I_k(x_1, \ldots, x_k)\, dx_k \approx$$

$$\approx \sum_{q_k=1}^{N_k} \frac{H_k}{2} \sum_{j=1}^{m_k} D_j^k I_k\Big(x_1, \ldots, x_{k-1}, \Big(q_k - \frac{1}{2}\Big) H_k + d_j^k \frac{H_k}{2}\Big). \qquad (1)$$

D'après (3.6.6) l'erreur dans cette formule est au plus

$$D(r_k) \max_{0 \leqslant x_k \leqslant 1} |(I_k(x_1, \ldots, x_k))_{x_k^{r_k}}|\, N_k^{-r_k}$$

et ne dépasse donc pas $D(r_k) A_k N_k^{-r_k}$. Pour une telle formule la somme des modules des coefficients des poids s'évalue par

$$\sum_{q_k=1}^{N_k} \frac{H_k}{2} \sum_{j=1}^{m_k} |D_j^k| = \frac{1}{2}\Big(\sum_{j=1}^{m_k}{}' |D_j^k|\Big) = B^k,$$

quantité qui ne dépend pas de N_k. Reprenant (5.6) il vient l'estimation du reste

$$R = E^0 + \sum_{q_1=1}^{N_1} D^{q_1} E_{q_1}^1 + \sum_{q_1=1}^{N_1} D^{q_1} \sum_{q_2=1}^{N_2} D^{q_2} E_{q_1 q_2}^2 + \ldots$$

D'où

$$|R| \leqslant |E^0| + B^1 \max |E_{q_1}^1| + B^1 B^2 \max |E_{q_1 q_2}^2| + \ldots \qquad (2)$$

Dans le cas considéré $E_{q_1 \ldots q_{k-1}}^{k-1}$ est l'erreur dans (1) pour certains $x_1, \ldots, x_{k-1}$ évalués au moyen de $D(r_k) A_k N_k^{-r_k}$. En substituant cette estimation dans le deuxième membre de (2), il vient

$$|R| \leqslant D_0(\mathbf{r}) \sum_{k=1}^{s} A_k N_k^{-r_k},$$

où D_0 (r) est défini par B^k et $D(r_k)$. L'estimation obtenue dépend du choix de formules de quadrature concrètes et du nombre de segments de division suivant les différents axes. Minimisons le facteur $\sum_{k=1}^{s} A_k N_k^{-r_k}$ en choisissant les valeurs de N_k pour un nombre fixe de points proportionnel à $N_1 \ldots N_s = Q$. Formons la fonction de Lagrange

$$\Phi = \sum_{k=1}^{s} A_k N_k^{-r_k} + \lambda (N_1 \ldots N_s - Q).$$

En égalant à zéro les dérivées Φ_{N_k}, on obtient le système d'équations

$$-A_k r_k N_k^{-r_k-1} + \lambda N_1 \ldots N_s N_k^{-1} = 0,$$

d'où

$$A_k r_k N_k^{-r_k} = \lambda N_1 \ldots N_s = \lambda Q, \qquad N_k = \left(\frac{A_k r_k}{\lambda Q}\right)^{1/r_k}.$$

L'équation $N_1 \ldots N_s = Q$ se met sous la forme

$$\prod_{k=1}^{s} \left(\frac{A_k r_k}{\lambda Q}\right)^{1/r_k} = Q$$

ou

$$(\lambda Q)^{1/r} = \frac{A}{Q}$$

avec

$$A = \prod_{k=1}^{s} (A_k r_k)^{1/r_k}, \quad \frac{1}{r} = \frac{1}{r_1} + \ldots \frac{1}{r_s};$$

il en découle $\lambda Q = \left(\frac{A}{Q}\right)^r$. Substituant sa valeur λQ dans l'expression pour N_k, il vient

$$N_k = \left(\frac{A_k r_k Q^r}{A^r}\right)^{1/r_k}.$$

On a alors

$$A_k N_k^{-r_k} = \frac{1}{r_k} \left(\frac{A}{Q}\right)^r,$$

et, par suite, la quantité $\sum_{k=1}^{s} A_k N_k^{-r_k}$ est $O\left(\left(\frac{A}{Q}\right)^r\right)$.

Considérons un cas particulier: $r_1 = \ldots = r_s = r_0$, $A_1 = \ldots = A_s = A_0$. Alors $\frac{1}{r} = \frac{1}{r_1} + \ldots + \frac{1}{r_s} = \frac{s}{r_0}$ et $A^r = A_0 r_0$.

Les formules considérées garantissent donc l'estimation d'erreur $O(A_0 N^{-r_0/s})$. En réalité, les dérivées bornées des fonctions sous le signe somme sont souvent d'ordre r_0 pas très grand ; c'est pourquoi pour des s grands, une telle évaluation de la vitesse de convergence se trouve mauvaise. On arrive parfois à améliorer l'ordre de l'estimation si, comme dans le cas unidimensionnel, on apprécie l'erreur au moyen des dérivées intégrables et non à l'aide des dérivées bornées. Cependant d'une façon générale, même des estimations pareilles ont la vitesse de convergence assez petite.

On voit se poser la question de formules de quadrature optimales sur les classes de fonctions à plusieurs variables. Comme ces formules ne sont connues pour aucune classe réelle de fonctions, nous nous bornerons à en minorer l'erreur.

§ 7. Minoration de l'erreur d'intégration numérique

Rappelons la position du problème d'optimisation des formules de quadrature sur une classe de fonctions. Supposons que la valeur approchée de l'intégrale se calcule au moyen de la formule

$$I(f) = \int_G f(P)\, p(P)\, dP \approx S_N(f) = \sum_{j=1}^{N} D_j f(P_j).$$

La quantité

$$R_N(f) = I(f) - S_N(f)$$

est appelée *erreur dans la formule de quadrature*, la quantité

$$R_N(F) = \sup_{f \in F} |R_N(f)|$$

erreur dans la formule de quadrature sur la classe F de fonctions et

$$W_N(F) = \inf_{D_j,\, P_j} R_N(F)$$

l'*estimation optimale de l'erreur dans les formules de quadrature sur la classe F* ; la formule (si elle existe) qui réalise cette borne inférieure est dite *optimale*. Nous supposerons accomplie la condition suivante : il existe un cube $\Delta \subset G$ dans lequel $p(P) \geqslant \gamma > 0$.

THÉORÈME.

$$W_N(C_{\mathbf{r}}(\mathbf{A})) \geqslant d(\Delta, \mathbf{r}, \mathbf{A})\, \gamma N^{-r},$$

avec $d(\Delta, \mathbf{r}, \mathbf{A}) > 0$.

On peut démontrer ce théorème de la façon suivante : pour tout ensemble de points $P_1, \ldots, P_N$ on peut construire une fonction

$f_{P_1, \ldots, P_N}(P)$ de la classe $C_r(A)$ considérée qui s'annule en tous ces points et telle que $I(f) \geqslant d(\Delta, r, A)\gamma N^{-r}$, la constante $d > 0$ ne dépendant pas des points $P_1, \ldots, P_N$. Alors, pour toute formule de quadrature avec ces points on a

$$R_N(C_r(A)) \geqslant |I(f_{P_1, \ldots, P_N}) - \sum_{j=1}^{N} D_j f_{P_1, \ldots, P_N}(P_j)| =$$
$$= |I(f_{P_1, \ldots, P_N})| \geqslant d(\Delta, r, A)\gamma N^{-r}.$$

Comme la quantité $R_N(C_r(A))$ est minorée par une constante ne dépendant pas des points P_j et des poids D_j, sa borne inférieure $W_N(C_r(A))$ sur l'ensemble de toutes les formules possibles est minorée par la même constante $d(\Delta, r, A)\gamma N^{-r}$. Donc, la démonstration du théorème se ramène à la construction de la fonction $f_{P_1, \ldots, P_N}(P)$ correspondante pour chaque ensemble de points $P_1, \ldots, P_N$.

Pour simplifier le calcul nous entreprendrons la construction pour le cas où Δ est le cube unité $0 \leqslant x_1, \ldots, x_s \leqslant 1$, $A_1 = \ldots = A_s = A_0$. Posons $n_k = [(2N)^{r/r_k}] + 1$ et partageons le cube Δ en $n_1 \ldots n_s$ parallélépipèdes rectangles $\Pi_m = \Pi_{m_1 \ldots m_s}$:

$$\frac{m_k - 1}{n_k} \leqslant x_k \leqslant \frac{m_k}{n_k}, \quad 0 < m_k \leqslant n_k, \quad k = 1, \ldots, s.$$

Soit $\varphi(x)$ une fonction indéfiniment dérivable sur la droite $(-\infty, \infty)$ telle que

$$\varphi(x) = 0 \quad \text{pour} \quad x \leqslant 0 \quad \text{et} \quad x \geqslant 1$$
$$0 < \varphi(x) \leqslant 1 \quad \text{pour} \quad 0 < x < 1.$$

Soit

$$\bar{r} = \max_{1 \leqslant k \leqslant s} r_k,$$
$$g = \max_{0 \leqslant l \leqslant \bar{r}} (2^l \max_{[0, 1]} |\varphi^{(l)}(x)|).$$

La définition des nombres n_k entraîne

$$(2N)^{r/r_k} \leqslant n_k \leqslant 2 \cdot (2N)^{r/r_k};$$

en multipliant ces inégalités et recourant à l'égalité $\frac{r}{r_1} + \ldots + \frac{r}{r_s} = 1$, on obtient

$$2N \leqslant n_1 \ldots n_s \leqslant 2^{s+1}N. \tag{1}$$

Formons la fonction

$$f_{P_1, \ldots, P_N}(P) = f_0(P)$$

de la façon suivante. Posons

$$f_0(P)=\frac{A_0}{g(2N)^r}\prod_{k=1}^{s}\varphi(n_kx_k-(m_k-1)) \tag{2}$$

dans les parallélépipèdes Π_m ne contenant à leur intérieur aucun des points $P_1, \ldots, P_s$. En tous les autres points posons $f_0(P)=0$. La fonction f_0 est indéfiniment dérivable à l'intérieur de chaque parallélépipède Π_m et s'annule avec toutes ses dérivées sur sa frontière. Donc, elle est indéfiniment dérivable dans tout l'espace $-\infty < x_1, \ldots, x_s < \infty$. Par construction, elle s'annule avec toutes ses dérivées en tous les points $P_1, \ldots, P_N$. Minorons $I(f_0)$. Utilisant l'inégalité $p(P) \geqslant \gamma$, après le changement de variables $x_k = \frac{m_k-1+y_k}{n_k}$, on obtient la chaîne de relations

$$\int_{\Pi_m} p(x_1, \ldots, x_s)\,A_0\,(g(2N)^r)^{-1}\prod_{k=1}^{s}\varphi(n_kx_k-(m_k-1))\,dx_1\ldots dx_s \geqslant$$

$$\geqslant\frac{A_0\gamma}{g(2N)^r n_1\ldots n_s}\int_0^1\ldots\int_0^1\prod_{k=1}^{s}\varphi(y_k)\,dy_1\ldots dy_s=\frac{d_1(r)\,A_0\gamma}{N^r n_1\ldots n_s}; \tag{3}$$

ici

$$d_1(r)=\frac{1}{g\cdot 2^r}\left(\int_0^1\varphi(y)\,dy\right)^s.$$

Chaque point P_j peut se trouver à l'intérieur d'un seul des parallélépipèdes Π_m. Donc, la fonction f_0 n'est pas identiquement nulle et est définie par l'égalité (2) dans $n_1\ldots n_s - N$ parallélépipèdes Π_m au moins. Puisque, en vertu de (1), on a $N \leqslant (n_1\ldots n_s)/2$, le nombre de ces parallélépipèdes est au moins égal à $(n_1\ldots n_s)/2$. Compte tenu de (3), on obtient l'estimation

$$I(f_0)\geqslant\frac{n_1\ldots n_s}{2}\,\frac{d_1(r)\,A_0\gamma}{N^r\cdot n_1\ldots n_s}=\frac{d(\Delta, r, A)\,\gamma}{N^r}, \tag{4}$$

où

$$d(\Delta, r, A)=\frac{d_1(r)\,A_0}{2}.$$

Puisque la fonction f_0 est indéfiniment dérivable, pour démontrer qu'elle appartient à la classe considérée il suffit d'évaluer ses dérivées en des points des parallélépipèdes en lesquels elle est différente

de zéro. En ces points on a

$$f x_j^{l_j} = \frac{A_0}{g\,(2N)^r}\,\varphi^{(l_j)}(n_j x_j - (m_j - 1))\,n_j^{l_j} \prod_{k \neq j} \varphi(n_k x_k - (m_k - 1)).$$

Comme

$$|\varphi| \leqslant 1,\ |\varphi^{(l_j)}| \leqslant g \cdot 2^{-l_j} \quad \text{pour} \quad l_j \leqslant \bar{r},$$

il en découle l'estimation

$$|f x_j^{l_j}| \leqslant \frac{A_0}{(2N)^r} \left(\frac{n_j}{2}\right)^{l_j}. \tag{5}$$

D'après la définition de n_j, on a

$$2 \leqslant n_j < 2\,(2N)^{r/r_j}.$$

Ainsi,

$$\left(\frac{n_j}{2}\right)^{l_j} \leqslant (2N)^{r l_j / r_j} \leqslant (2N)^r$$

pour $0 \leqslant l_j \leqslant r_j$. Utilisant (5) on obtient l'estimation $|f x_j^{l_j}| \leqslant$ $\leqslant A_0$ pour $0 \leqslant l_j \leqslant r_j$. Donc, la fonction construite appartient à la classe considérée, vérifie l'inégalité (4) où la constante d (Δ, r, A) ne dépend pas des points d'intégration et s'annule en tous les points P_j; ainsi nous avons défini un procédé de construction de la fonction cherchée pour tout ensemble de points $P_1, \ldots, P_N$.

REMARQUE. La fonction construite s'annule aux points $P_1, \ldots$ $\ldots, P_N$ avec toutes ses dérivées. C'est pourquoi nous aurions obtenu la même minoration pour la classe de toutes les formules de quadrature possibles à partir des valeurs de la fonction et de ses dérivées en N points.

§ 8. Optimisation de l'estimation de l'erreur sur des classes plus vastes de procédés d'intégration

Au § 7 nous avons minoré l'erreur dans les formules de quadrature, c'est-à-dire les formules d'intégration

$$I(f) \approx \sum_{j=1}^{N} D_j f(P_j),$$

les points P_j et les poids D_j ne dépendant pas de la fonction à intégrer. Pour nombre de problèmes, on préfère, en raison de leur plus grande efficacité, les procédés d'intégration où le choix des points suivants dépend de l'information sur les valeurs déjà calculées de

la fonction, par exemple, les procédés avec une recherche automatique du pas (§§ 17, 18, ch. III).

Soit un procédé d'intégration tel qu'on ne retienne de l'information sur la fonction sous le signe somme que celle sur ses valeurs en des points isolés. Ce procédé est défini en se donnant le premier point d'intégration, la règle utilisée pour trouver les points suivants et la technique de calcul de la valeur approchée de l'intégrale. Ainsi, chacun de ces procédés se ramène au schéma suivant : on se donne un point P_1 et les fonctions

$$Q_q = \Phi_q (Q_1, \ldots, Q_{q-1}; y_1, \ldots, y_{q-1}), \quad q = 2, \ldots, N,$$
$$S_N (Q_1, \ldots, Q_N; y_1, \ldots, y_N),$$

Q_j étant des points du domaine G, y_j, S_N des nombres. Dans le calcul approché d'une intégrale concrète on trouve successivement les valeurs

$$f(P_1), \; P_2 = \Phi_2 (P_1; f(P_1)), \; f(P_2),$$
$$P_3 = \Phi_3 (P_1, P_2; f(P_1), f(P_2)),$$
$$f(P_3), \quad \ldots, \quad f(P_N)$$

et on pose ensuite

$$I(f) \approx S_N (P_1, \ldots, P_N; f(P_1), \ldots, f(P_N)).$$

Le point P_1 étant donné en même temps que les fonctions Φ_j et S_N, la dépendance de celles-ci par rapport à P_1 pourrait être omise.

Comme dans le cas des formules de quadrature, on peut déterminer l'erreur de méthode dans le calcul de l'intégrale donnée :

$$\tilde{R}_N (f) = I(f) - S_N (P_1, \ldots, P_N; f(P_1), \ldots, f(P_N)),$$

l'erreur de méthode sur la classe des fonctions sous le signe somme

$$\tilde{R}_N (F) = \sup_{f \in F} |\tilde{R}_N (f)|$$

et l'estimation optimale de l'erreur sur l'ensemble de tous les procédés d'intégration possibles

$$Z_N (F) = \inf_{P_1, \Phi_2, \ldots, \Phi_N, S_N} \tilde{R}_N (F).$$

THÉORÈME. *Supposons que la classe F de fonctions est un compact convexe à symétrie centrale de centre $f \equiv 0$ et que toutes les fonctions de cette classe sont uniformément bornées. Alors les estimations optimales de l'erreur sur les ensembles de toutes les formules de quadrature et de tous les procédés d'intégration possibles coïncident :*

$$W_N (F) = Z_N (F).$$

La condition de convexité de la classe F signifie que toutes les fois qu'elle contient f_1, f_2 quelconques, elle contient également toutes

les fonctions $f_1\theta + f_2(1-\theta)$, où $0 < \theta < 1$. La condition que la classe jouit de la symétrie centrale de centre $f \equiv 0$ veut dire que toutes les fois qu'elle contient toute fonction f, elle contient également la fonction $-f$. En particulier, toutes les classes C_r (A) satisfont aux conditions du théorème.

Démontrons en bref ce théorème. Prenons $\eta > 0$ arbitraire. Conformément à la définition de $Z_N(F)$, on peut indiquer un procédé d'intégration tel que

$$\widetilde{R}_N(F) \leqslant Z_N(F) + \eta.$$

Pour $P_1, \Phi_2, \ldots, \Phi_N$ correspondant à ce procédé posons

$$P_1^0 = P_1, \quad P_2^0 = \Phi_2(P_1^0; 0), \ \ldots,$$

$$P_N^0 = \Phi_N(P_1^0, \ \ldots, \ P_{N-1}^0; \ 0, \ \ldots, \ 0).$$

Désignons par Y l'ensemble des points d'un espace à $(N+1)$ dimensions

$$(y_0, y_1, \ldots, y_N) = (I(f), f(P_1^0), \ldots, f(P_N^0))$$

pour $f \in F$. D'après les propriétés de la classe F, l'ensemble Y est symétrique,

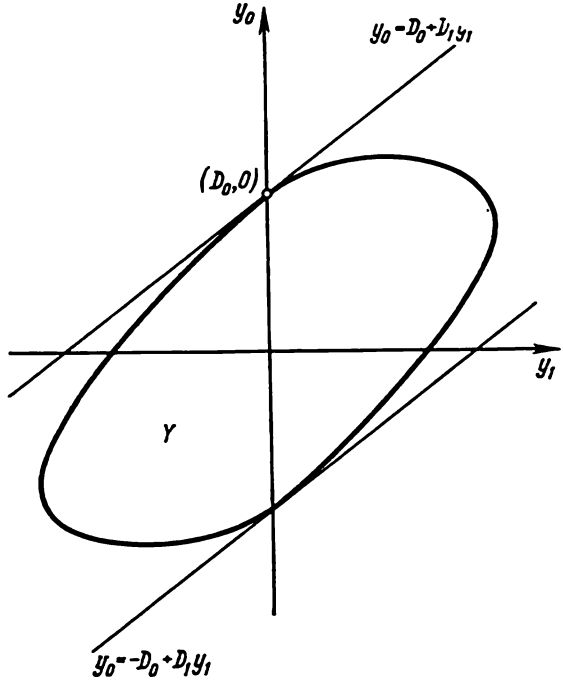


Fig. 5.8.1

de centre de symétrie à l'origine des coordonnées, fermé et contenu dans une partie finie de l'espace (fig. 5.8.1). Soit

$$D_0 = \sup_{(y_0, \ldots, y_N) \in Y} y_0.$$

Puisque Y est fermé, le point $(D_0, 0, \ldots, 0)$ appartient à Y et il existe une fonction $f_0 \in F$ qui lui correspond et telle que

$$I(f_0) = D_0,\ f_0(P_1^0) = \ldots = f_0(P_N^0) = 0.$$

La théorie des ensembles convexes nous apprend que par tout point de la frontière d'un ensemble on peut mener un plan d'appui; rappelons qu'un plan d'appui est un plan passant par un point de la frontière et tel que tous les points de l'ensemble soient d'un seul côté de ce plan. Soit

$$\alpha_0(y_0 - D_0) + \sum_{j=1}^{N} \alpha_j y_j = 0$$

l'équation de ce plan d'appui. Nous proposons au lecteur de s'assurer que parmi tous les plans d'appui passant par le point $(D_0, 0, \ldots, 0)$ il en existe un qui vérifie $\alpha_0 \neq 0$. Son équation se met sous la forme

$$y_0 = D_0 + \sum_{j=1}^{N} D_j y_j, \quad D_j = -\frac{\alpha_j}{\alpha_0}.$$

Le plan qui lui est symétrique par rapport à l'origine $(0, \ldots, 0)$ est

$$y_0 = -D_0 + \sum_{j=1}^{N} D_j y_j.$$

Par suite de la symétrie du corps Y par rapport à l'origine, ce plan sera également un plan d'appui pour Y. Le corps Y se trouve entre ces plans; donc, pour tous ses points

$$\left| y_0 - \sum_{j=1}^{N} D_j y_j \right| \leqslant D_0;$$

se rappelant la signification des coordonnées y_j, écrivons cette relation sous la forme

$$\left| I(f) - \sum_{j=1}^{N} D_j f(P_j^0) \right| \leqslant D_0$$

pour tout $f \in F$.

Nous avons obtenu une formule de quadrature avec estimation du reste $R_N(F) \leqslant D_0$. Donc, l'inégalité

$$W_N(F) = \inf R_N(F) \leqslant D_0$$

est satisfaite. Appliquons l'algorithme initial au calcul des intégrales $I(\pm f_0)$. Vu la condition $f_0(P_j^0) = 0$, j quelconque, nous obtiendrons successivement les valeurs des points et celles des fonctions en ces points:

$$\pm f_0(P_1^0) = 0, \quad P_2 = \Phi_2(P_1^0;\ \pm f(P_1^0)) =$$

$$= \Phi_2(P_1^0;\ 0) = P_2^0 \pm f_0(P_2^0) = 0, \ \ldots, \ \pm f_0(P_{N-1}^0) = 0,$$

$$P_N = \Phi_N(P_1^0, \ldots, P_{N-1}^0;\ \pm f_0(P_1^0), \ \ldots, \ \pm f_0(P_{N-1}^0)) =$$

$$= \Phi_N(P_1^0, \ \ldots, \ P_{N-1}^0;\ 0, \ \ldots, \ 0) = P_N^0, \ \pm f_0(P_N^0) = 0.$$

Dans les deux cas, la valeur approchée de l'intégrale sera posée égale à

$$S^0 = S(P_1^0, \ldots, P_N^0; \pm f_0(P_1^0), \ldots, \pm f_0(P_N^0)) = S(P_1^0, \ldots, P_N^0; 0, \ldots, 0).$$

D'autre part,

$$I(\pm f_0) = \pm D_0,$$

aussi

$$|I(f_0) - S^0| + |S^0 - I(-f_0)| \geqslant |(I(f_0) - S^0) + (S^0 - I(-f_0))| = 2D_0.$$

Ainsi, pour l'une des fonctions $\pm f_0$ l'erreur sur la valeur approchée est au moins égale à D_0. En d'autres termes, $D_0 \leqslant \widetilde{R}_N(F)$. On a la chaîne des relations

$$W_N(F) \leqslant R_N(F) \leqslant D_0 \leqslant \widetilde{R}_N(F) \leqslant Z_N(F) + \eta,$$

d'où, η étant arbitraire,

$$W_N(F) \leqslant Z_N(F).$$

L'ensemble de tous les procédés d'intégration possibles contient l'ensemble des formules de quadrature. C'est pourquoi la borne inférieure de l'erreur sur le premier ensemble est au plus égale à la borne inférieure de l'erreur sur le second

$$Z_n(F) \leqslant W_N(F).$$

Donc

$$W_N(F) = Z_N(F).$$

Les classes de fonctions qui interviennent dans les divers chapitres de la théorie de l'intégration numérique satisfont toujours aux conditions du théorème démontré ; on pourrait en déduire que l'étude des procédés d'intégration plus généraux n'est pas plus avantageuse que celle des formules de quadrature et n'est donc pas rationnelle. Du point de vue d'une science appliquée cette conclusion est fausse. Citons les arguments à l'appui.

1. En pratique les classes de fonctions peuvent tout simplement ne pas être convexes. Une classe typique que l'on rencontre dans les applications est visiblement celle des fonctions analytiques par morceaux. Reprenons le cas unidimensionnel. Si la classe est convexe, la demi-somme de deux fonctions de la classe appartient toujours à cette classe. La demi-somme des fonctions présentant l points où l'analyticité n'a plus lieu peut avoir $2l$ points jouissant de cette propriété. Donc, la classe des fonctions analytiques telles qu'elles ne le sont plus en l points donnés n'est pas convexe. Si le nombre de tels points n'est pas majoré, la classe n'est pas fermée : la limite d'une suite de fonctions analytiques par morceaux telles que le nombre de points où elles ne sont pas analytiques croît indéfiniment peut être une fonction qui n'est pas analytique par morceaux.

2. Plus haut on a comparé les méthodes d'intégration d'après la borne supérieure de l'erreur sur une classe. Il se peut cependant que les deux méthodes présentent une même erreur sur la classe, tandis que pour la plupart des fonctions de celle-ci l'une d'elles a une erreur plus petite. Il est clair que cette dernière méthode est à préférer, et la comparaison des méthodes selon la borne supérieure de l'erreur sur une classe ne donne pas dans ce cas un tableau d'ensemble.

§ 9. Méthode de Monte-Carlo

En construisant des formules de quadrature on a toujours évalué leur erreur sur une certaine classe de fonctions. Pour la formule des trapèzes par exemple c'était une estimation de la forme const $\cdot A_2 N^{-2}$ sur la classe des fonctions ayant leurs dérivées secondes bornées par A_2 constant (N est le nombre de points d'intégration). Il s'agit en l'occurrence d'*estimations garanties de l'erreur sur une classe de fonctions*, i.e. elles garantissent que l'erreur sur la valeur approchée de l'intégrale ne dépasse pas une certaine valeur pour toutes les fonctions sous le signe somme appartenant à la classe considérée. Dans cette approche, en évaluant l'erreur commise sur une intégrale concrète, on se guide sur l'erreur d'intégration de la « pire » fonction de la classe en question. Pour certaines classes cette évaluation est si mauvaise qu'elle ôte tout espoir d'obtenir une valeur approchée de l'intégrale avec la précision demandée. Par exemple, d'après les théorèmes des §§ 7, 8, il n'existe pas de méthodes dans lesquelles l'estimation de l'erreur sur la classe des fonctions $C_{1,\ldots,1}(A, \ldots, A)$ soit meilleure que $dAN^{-1/s}$ (s étant la multiplicité de l'intégrale calculée).

Supposons qu'on a à garantir une estimation inférieure à 0,01 dA. Alors le nombre de points doit vérifier l'inégalité $dAN^{-1/s} \leqslant 0{,}01\, dA$, autrement dit, $N \geqslant 100^s$. Puisque le calcul de chaque valeur de la fonction sous le signe somme exige en général un grand nombre d'opérations arithmétiques, même pour $s = 6$ une telle condition imposée au nombre de points se trouve pratiquement impossible à réaliser.

Ainsi, il nous est impossible de calculer la valeur de l'intégrale avec une estimation garantie d'erreur égale à 0,01 dA. On lève cette contradiction en décrivant de façon plus détaillée les classes de fonctions sous le signe somme. En particulier, nous laissons de côté les classes de fonctions et les méthodes d'intégration numérique correspondantes qui sont apparues lors du calcul d'intégrales résultant de la résolution d'équations par développement en série de Neumann. On peut aussi refuser une estimation stricte, garantie de l'erreur au profit d'une autre qui n'a qu'un certain degré de certitude. En particulier, en confectionnant les programmes standard d'intégration (§ 17, ch. III) nous avons renoncé à évaluer l'erreur de façon stricte ; celle-ci a été évaluée au moyen de la différence des résultats du calcul de la valeur approchée de l'intégrale pour diverses valeurs des paramètres.

Une des méthodes de calcul approché des valeurs d'intégrales telle que l'erreur est évaluée de façon non garantie mais avec un certain degré de certitude est la *méthode de Monte-Carlo*.

Soit à calculer la valeur approchée de l'intégrale

$$I(f) = \int_G f(P)\, dP.$$

Supposons qu'on a réussi par un procédé quelconque à obtenir N points aléatoires $P_1, \ldots, P_N$ deux à deux indépendants, équirépartis dans G avec la densité $p(P)$, $\int_G p(P)\,dP = 1$. Posons

$$s_j(f) = \frac{f(P_j)}{p(P_j)}.$$

Les variables aléatoires $s_j(f)$ sont deux à deux indépendantes et possèdent une même loi:

$$\mathrm{E}s_j(f) = \int_G \frac{f(P)}{p(P)}\, p(P)\,dP = I(f)$$

et

$$\mathrm{Var}(s_j(f)) = \mathrm{E}s_j^2(f) - (\mathrm{E}s_j(f))^2 =$$

$$= \int_G \left(\frac{f(P)}{p(P)}\right)^2 p(P)\,dP - (\mathrm{E}s_j(f))^2 = D(f),$$

où

$$D(f) = I\left(\frac{f^2}{p}\right) - (I(f))^2.$$

Posons

$$S_N(f) = \frac{1}{N}\sum_{j=1}^{N} s_j(f).$$

D'après les propriétés indiquées des variables $s_j(f)$ on a

$$\mathrm{E}S_N(f) = \frac{1}{N}\sum_{j=1}^{N} \mathrm{E}s_j(f) = I(f),$$

$$\mathrm{Var}(S_N(f)) = \frac{1}{N^2}\sum_{j=1}^{N} \mathrm{Var}(s_j(f)) = \frac{1}{N} D(f).$$

L'inégalité

$$|S_N(f) - I(f)| \leqslant \sqrt{\frac{D(f)}{\eta N}} \tag{1}$$

(dite *inégalité de Tchebycheff*) est satisfaite avec la probabilité $1-\eta$. En faisant $\eta = 0{,}01$, il vient

$$|S_N(f) - I(f)| \leqslant 10\sqrt{\frac{D(f)}{N}}$$

avec la probabilité 99 %.

On obtiendrait une meilleure estimation en supposant que les points P_j sont non seulement deux à deux indépendants mais indépendants dans leur ensemble. Alors, d'après le théorème limite central la variable aléatoire

$$\frac{S_N(f)-I(f)}{\sqrt{D(f)/N}}$$

a une répartition asymptotiquement normale et la fonction de répartition

$$\Phi(y)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{y}\exp\left(-\frac{t^2}{2}\right)dt.$$

La probabilité pour que la variable aléatoire ayant une telle fonction de répartition ne dépasse pas en module la valeur $y>0$ est asymptotiquement égale à

$$p_0(y)=1-\frac{1}{\sqrt{\frac{2}{\pi}}}\int_{y}^{\infty}\exp\left(-\frac{t^2}{2}\right)dt.$$

Ainsi, pour N grands on a, avec une probabilité voisine de $p_0(y)$, l'inégalité

$$|S_N(f)-I(f)|\leqslant y\sqrt{\frac{D(f)}{N}}.$$

En posant $y=3$ et $y=5$, on obtient que les inégalités

$$|S_N(f)-I(f)|\leqslant 3\sqrt{\frac{D(f)}{N}} \quad \text{et} \quad |S_N(f)-I(f)|\leqslant 5\sqrt{\frac{D(f)}{N}}$$

sont satisfaites avec les probabilités respectives 0,997 et 0,99999. Les assertions énoncées ci-dessus sont parfois appelées les règles « des trois sigma » et « des cinq sigma » respectivement.

La quantité inconnue

$$D(f)=I\left(\frac{f^2}{p}\right)-(I(f))^2$$

intervient dans le deuxième membre de ces estimations. Tâchons de l'évaluer à partir de l'information sur les valeurs calculées de $f(P_j)$. D'après la définition de $D(f)$, on a

$$D(f)=\mathrm{E}\,(s_j(f)-I(f))^2.$$

Puisque $S_N(f)\approx I(f)$, on a $D(f)\approx l(f)$, où

$$l(f)=\mathrm{E}\,(s_j(f)-S_N(f))^2.$$

Recourons à des raisonnements plus rigoureux.

La relation

$$\mathrm{E}\,(s_j(f) - S_N(f)) = \mathrm{E}s_j(f) - \mathrm{E}S_N(f) = I(f) - I(f) = 0$$

entraîne $l(f) = \mathrm{Var}\,(s_j(f) - S_N(f))$. On a

$$s_j(f) - S_N(f) = s_j(f) - \frac{1}{N}\sum_{i=1}^{N} s_i(f) =$$

$$= s_j(f)\left(1 - \frac{1}{N}\right) - \frac{1}{N}\sum_{i\neq j} s_i(f);$$

il en résulte

$$\mathrm{Var}\,(s_j(f) - S_N(f)) = \left(1 - \frac{1}{N}\right)^2 \mathrm{Var}\,(s_j(f)) +$$

$$+ \sum_{i\neq j} \mathrm{Var}\,(s_i(f))/N^2 = \left(\left(1 - \frac{1}{N}\right)^2 + \frac{N-1}{N^2}\right) D(f) = \left(1 - \frac{1}{N}\right) D(f).$$

On a définitivement $l(f) = \left(1 - \frac{1}{N}\right) D(f)$ et, donc,

$$D(f) = \frac{N}{N-1}\,\mathrm{E}\,(s_j(f) - S_N(f))^2.$$

Selon la loi des grands nombres on peut démontrer que l'égalité approchée

$$\mathrm{E}\,(s_j(f) - S_N(f))^2 \approx \frac{1}{N}\sum_{j=1}^{N} (s_j(f) - S_N(f))^2$$

est satisfaite avec une grande probabilité. La dernière relation entraîne avec une grande probabilité l'égalité approchée

$$D(f) \approx D^{(N)}(f),$$

où

$$D^{(N)}(f) = \frac{1}{N-1}\sum_{j=1}^{N} (s_j(f) - S_N(f))^2.$$

Citons à titre d'exemple un schéma d'application de la méthode de Monte-Carlo. Considérons le cas $\mu(G) = 1$. Soient $1 - \eta$ le niveau de sécurité avec lequel on veut obtenir la valeur approchée de l'intégrale, E la précision donnée. On détermine y par l'égalité

$$\eta = \frac{1}{\sqrt{\pi}} \int_0^{\infty} \exp\left(-\frac{t^2}{2}\right) dt.$$

Pour $n = 1, \ldots$ on obtient successivement les points aléatoires P_n indépendants, uniformément répartis dans G et on calcule $t_n(f)$, $S_n(f)$, $d_n(f)$, $D_n(f)$ en utilisant les relations de récurrence

$$t_n(f) = t_{n-1}(f) + f(P_n),$$

$$S_n(f) = \frac{t_n(f)}{n},$$

$$d_n(f) = d_{n-1}(f) + \frac{n}{n-1}(f(P_n) - S_n(f))^2,$$

$$D_n(f) = \frac{d_n(f)}{n-1},$$

et la quantité

$$\lambda_n = y\sqrt{\frac{D_n(f)}{n}}.$$

Les conditions initiales de la récursion sont

$$t_1(f) = S_1(f) = f(P_1),\ d_1(f) = D_1(f) = 0.$$

PROBLÈME. Vérifier que $D_n(f) = D^{(n)}(f)$.

Si $\lambda_n \leqslant E$, les calculs s'arrêtent; on pose $I(f) \approx S_N(f)$ et on considère que l'inégalité $|I(f) - S_N(f)| \leqslant E$ est satisfaite avec la probabilité $1-\eta$.

Remarquons qu'en réalité cette inégalité est satisfaite avec une probabilité un peu moindre, mais très proche de $1 - \eta$.

On lit souvent dans les manuels de calcul numérique que la méthode de Monte-Carlo est une technique universelle pour calculer les intégrales de multiplicité élevée. L'ordre de grandeur de l'estimation de la vitesse de convergence de la méthode de Monte-Carlo, $O(1/\sqrt{N})$, ne dépend pas de la dimension de l'intégrale, alors que celui des estimations garanties augmente sensiblement avec la dimension. Dans la méthode de Monte-Carlo, pour tout n, on évalue effectivement l'erreur au moyen de $\sqrt{D_n(f)/n}$ et, donc, le calcul s'arrête au moment où la précision demandée est atteinte.

Avant d'y souscrire voyons le « pour » et le « contre » de la méthode en question.

§ 10. Légitimité des méthodes indéterministes de résolution des problèmes

Certains utilisateurs sont prévenus contre la méthode de Monte-Carlo : la petitesse de son erreur n'étant assurée qu'avec une certaine probabilité, ils l'avancent pour nier sa légitimité.

Nous avons déjà dit qu'en calculant les intégrales de multiplicité très élevée on n'a aucun espoir (ou peu s'en faut) d'obtenir une valeur

approchée de l'intégrale avec la garantie d'une petite estimation d'erreur. Cette situation suggère justement l'emploi de la méthode de Monte-Carlo.

Remarquons que même les programmes standard de calcul des intégrales simples (ch. III) ne garantissent pas complètement une erreur faible en cas de fonctions à intégrer régulières arbitraires. D'autre part, la résolution de n'importe quel problème s'accompagne d'erreurs éventuelles dans sa position, dans le programme, etc. Pour toutes ces raisons et pour d'autres encore il est rare qu'on puisse assurer que l'écart entre le résultat des calculs et le modèle réel sera petit : il y a toujours une certaine probabilité de l'erreur. Tout cela nous montre que l'abandon de la méthode de Monte-Carlo à cause de son caractère probabiliste est injustifié.

D'autre part, en utilisant la méthode de Monte-Carlo on doit tenir compte de ses défauts suivants :

Pour appliquer la méthode de Monte-Carlo il est nécessaire de disposer d'une suite de points indépendants P_j possédant une loi de répartition donnée. En pratique, l'utilisateur dispose des générateurs de nombres aléatoires ou pseudo-aléatoires qui donnent des suites de nombres aléatoires uniformément répartis sur [0, 1]. En les transformant on obtient des nombres aléatoires de lois demandées. Les premières calculatrices électroniques étaient munies de certains dispositifs (utilisant par exemple le phénomène de désintégration radioactive) qui élaboraient des suites de variables aléatoires satisfaisant parfois à la condition d'indépendance dans l'ensemble. Or, leur vitesse de fonctionnement était petite et on y renonça avec le progrès du matériel de calcul. Au lieu de générateurs de nombres aléatoires on utilise des générateurs de nombres pseudo-aléatoires : programmes qui donnent les suites des nombres à prendre pour nombres aléatoires. L'utilisation des générateurs de nombres pseudo-aléatoires s'est avérée bien avantageuse car elle a permis une vaste application des méthodes probabilistes. Pourtant, on doit préciser toujours les propriétés des suites de nombres qu'ils fournissent. En effet, certains générateurs élaborent des suites de nombres qui peuvent être considérés comme deux à deux indépendants, mais non indépendants dans leur ensemble. Dans ce cas on ne saurait se servir d'estimations d'erreur basées sur le théorème limite central. On désire calculer l'intégrale

$$I(f) = \int_0^1 \dots \int_0^1 f(x_1, \dots, x_s)\, dx_1 \dots dx_s$$

par la méthode de Monte-Carlo. Supposons qu'on veut prendre comme points d'intégration une suite de points indépendants uniformément répartis du cube unité. Si le générateur de nombres pseudo-aléatoires donne une suite de nombres $\xi_1, \dots$ uniformément répartis

sur [0, 1], il est naturel de prendre pour points d'intégration les points $(\xi_1, \ldots, \xi_s)$, $(\xi_{s+1}, \ldots, \xi_{2s})$, $\ldots$

Pour que l'inégalité de Tchebycheff soit licite, il faut que l'hypothèse sur l'indépendance de la répartition de tous les points P_j, c'est-à-dire l'indépendance de la répartition des ensembles

$$(\xi_{(j-1)s+1}, \ldots, \xi_{js}), \quad (\xi_{(l-1)s+1}, \ldots, \xi_{ls}),$$

soit remplie.

A mesure que s croît, cette hypothèse impose des contraintes de plus en plus sévères au générateur de nombres pseudo-aléatoires. On connaît de nombreux exemples d'une utilisation malheureuse de la méthode de Monte-Carlo pour s grands, dont voici la raison. La mise en œuvre de la méthode et l'évaluation d'erreur se faisaient dans certaines hypothèses sur telles ou telles propriétés statistiques des nombres pseudo-aléatoires. Or ces hypothèses étaient fausses. Ce qui entraîna une fausse appréciation de la petitesse de l'erreur.

Donc, le danger de la méthode de Monte-Carlo n'est pas tant dans le caractère probabiliste de l'évaluation de l'erreur que dans le fait qu'on effectue souvent celle-ci dans l'hypothèse des propriétés inexistantes des générateurs de nombres pseudo-aléatoires.

L'indépendance de l'ordre de grandeur de l'estimation d'erreur par rapport à la dimension de l'intégrale qu'on calcule constitue un avantage de la méthode considérée. Pourtant, en insistant sur l'ordre de convergence de la méthode on risque de négliger le fait important suivant. Nous avons évalué l'erreur sur la valeur approchée de l'intégrale et obtenu

$$|S_N(f) - I(f)| \leqslant \text{const} \cdot \sqrt{D(f)/N}.$$

Or d'habitude, c'est la petitesse de l'erreur relative sur la valeur approchée de l'intégrale qui importe, i.e., dans notre cas, celle de la quantité

$$\frac{\sqrt{D(f)}}{|I(f)|\sqrt{N}}.$$

La statistique des intégrales qu'on calcule en réalité montre que $\sqrt{D(f)}/|I(f)|$ a tendance à s'accroître brusquement avec la dimension des intégrales. Illustrons ce fait sur l'exemple de l'intégrale

$$I(f_s) = \int_0^1 \ldots \int_0^1 \exp\left(-32\left(x_1^2 + \ldots + x_s^2\right)\right) dx_1 \ldots dx_s,$$

telle que

$$\frac{\sqrt{D(f_s)}}{|I(f_s)|} > 10^{s/2} - 1.$$

De ce fait le coût de la méthode de Monte-Carlo augmente sensiblement avec la dimension des intégrales (l'erreur relative étant la même). Avant d'appliquer directement la méthode au calcul d'intégrales multiples on se propose souvent de diminuer la quantité $\sqrt{D(f)}/|I(f)|$ et on procède à cet effet à une étude minutieuse des propriétés de la fonction sous le signe somme, à la transformation de l'intégrale moyennant un changement de variables, etc.

§ 11. Accélération de la convergence de la méthode de Monte-Carlo

Voyons certains procédés susceptibles d'augmenter l'efficacité de la méthode de Monte-Carlo.

1. Il s'agit en particulier de représenter $f(P)$ sous la forme

$$f(P) = F(P) + g(P),$$

où $F(P)$ est explicitement intégrable et contient toutes les composantes à variation rapide de $f(P)$, et $g(P)$ est une fonction variant continûment de variance petite. On divise parfois le domaine d'intégration en petits domaines partiels et on prend dans chacun de ceux-ci pour $F(P)$ un polynôme d'interpolation aux points dans ce domaine partiel.

2. La variante de la méthode de Monte-Carlo étudiée au § 9 comportait comme paramètre libre la fonction $p(P)$, densité de répartition des points d'intégration. L'estimation d'erreur montre que l'erreur est d'autant plus petite que les valeurs de la fonction $f(P)/p(P)$ sont moins dispersées. Un choix convenable de la fonction $p(P)$ entraîne donc également une diminution de la variance. Nous avons étudié le cas où tous les points P_j avaient une même fonction de répartition $p(P)$; il s'avère souvent utile de choisir des points d'intégration ayant des fonctions de répartition $p_j(P)$ distinctes.

Un cas particulier répandu de ce procédé est le suivant. L'intégrale initiale est représentée par la somme d'intégrales

$$I(f) = \int_G f(P)\,dP = \sum_{l=1}^{m} I_l(f), \quad I_l(f) = \int_{G_l} f(P)\,dP,$$

N par

$$N = N_1 + \ldots + N_m,$$

et on calcule chaque $I_l(f)$ par la méthode de Monte-Carlo à N_l points d'intégration. Etudions le cas de l'intégrale de la fonction visualisée par la fig. 5.11.1. En calculant directement l'intégrale initiale pour $p \equiv 1$ on a $D(f) = 1/4$. Si $G_1 = [0, 1/2]$, $G_2 = [1/2, 1]$, alors pour tout N_1, $N_2 \neq 0$, les deux intégrales $I_l(f)$ et, partant,

l'intégrale initiale, seront calculées exactement. Ce cas très rare ne se produit que si on a réussi à partager le domaine initial en parties dans lesquelles la fonction sous le signe somme est constante; pourtant, si on parvient à le partager en des parties telles que la fonction y varie faiblement, on peut obtenir une amélioration sensible de la précision pour le même volume de calcul.

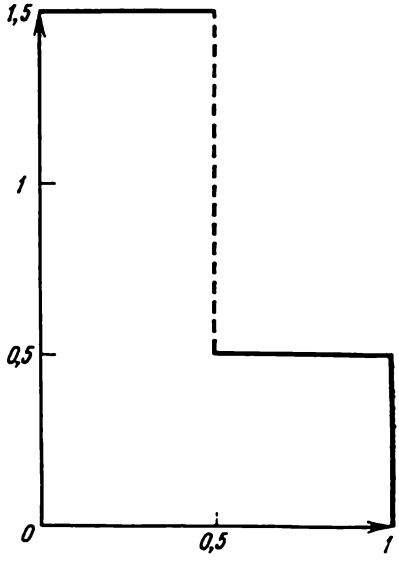


Fig. 5.11.1

Considérons le cas où le calcul de chaque intégrale $I_l(f)$ se fait avec les points P_j^l, $j = 1, \ldots, N_l$, dont la densité de répartition $p_l = \mu_l^{-1}$ est uniforme dans le domaine G_l, et tels que deux points quelconques (appartenant à un même domaine ou à des domaines différents) sont répartis indépendamment; ici μ_l désigne la mesure du domaine G_l. Comparons la méthode décrite avec le schéma initial de la méthode de Monte-Carlo dans lequel les points d'intégration sont répartis uniformément et deux à deux indépendants. Alors on a pour ce schéma

$$p(P) = \mu^{-1}, \quad \operatorname{Var} S_N(f) = D(f) N^{-1},$$

où μ est la mesure du domaine G,

$$D(f) = \mu I(f^2) - (I(f))^2.$$

On prend pour valeurs approchées des intégrales $I_l(f)$ les quantités

$$S^l(f) = \frac{\mu_l}{N_l} \sum_{j=1}^{N_l} f(P_j^l).$$

On a de même

$$\mathrm{E}S^l(f) = I_l(f),$$
$$\operatorname{Var} S^l(f) = D_l(f) N_l^{-1},$$

où

$$D_l(f) = \mu_l I_l(f^2) - (I_l(f))^2.$$

Posons

$$\overline{S}_N(f) = \sum_{l=1}^{m} S^l(f);$$

si deux points P_n^i et P_q^j quelconques sont répartis de façon indépendante, les quantités $S^l(f)$ sont également deux à deux indépendantes, donc,

$$\mathrm{E}\overline{S}_N(f) = \sum_{l=1}^{m} \mathrm{E}S^l(f) = I(f),$$

$$\mathrm{Var}\,\overline{S}_N(f) = \sum_{l=1}^{m} \mathrm{Var}\,(S^l(f)) = \sum_{l=1}^{m} D_l(f)\,N_l^{-1}.$$

Considérons le problème de minimisation de $\mathrm{Var}\,(\overline{S}_N(f))$ par le choix des valeurs $N_1, \ldots, N_m$, le nombre total des points $N = N_1 + N_2 + \ldots + N_m$ étant fixe. Introduisons la fonction de Lagrange

$$\Psi = \mathrm{Var}\,\overline{S}_N(f) - \lambda\left(\sum_{l=1}^{m} N_l - N\right);$$

en égalant à zéro ses dérivées par rapport à N_l, on obtient le système d'équations $-D_l(f)\,N_l^{-2} - \lambda = 0$; en résolvant ce système simultanément avec l'équation $N = N_1 + \ldots + N_m$, on détermine les valeurs demandées N_l, Ici et plus bas, comme d'habitude, on néglige le fait que les valeurs N_l doivent être entières.

Montrons que pour $N_l = N\mu_l\mu^{-1}$ on a l'inégalité

$$\mathrm{Var}\,(\overline{S}_N(f)) \leqslant \mathrm{Var}\,(S_N(f)).$$

LEMME. *Soit ξ une variable aléatoire. Pour tout a on a alors*

$$\mathrm{E}(\xi - a)^2 \geqslant \mathrm{E}(\xi - \mathrm{E}\xi)^2 = \mathrm{Var}\,\xi.$$

Nous laissons au lecteur le soin de démontrer ce lemme.

Pour $N_l = N\mu_l\mu^{-1}$ on peut écrire la chaîne de relations

$$\mathrm{Var}\,S_N(f) = \frac{D(f)}{N} = \frac{1}{N}\int_G \frac{1}{\mu}(\mu f(P) - I(f))^2\,dP =$$

$$= \sum_{l=1}^{m} \frac{\mu}{N}\int_{G_l}\left(f(P) - \frac{I(f)}{\mu}\right)^2 dP =$$

$$= \sum_{l=1}^{m} \frac{\mu}{N\mu_l}\int_{G_l}\mu_l\left(f(P) - \frac{I(f)}{\mu}\right)^2 dP =$$

$$= \sum_{l=1}^{m} \frac{1}{N_l}\int_{G_l}\frac{1}{\mu_l}\left(\mu_l f(P) - \frac{\mu_l}{\mu}I(f)\right)^2 dP =$$

$$= \sum_{l=1}^{m} \frac{1}{N_l}\mathrm{E}_l\left(\mu_l f(P) - \frac{\mu_l}{\mu}I(f)\right)^2; \qquad (1)$$

on désigne ici par $E_l g(P)$ l'espérance mathématique de $g(P)$ sous l'hypothèse que le point P est uniformément réparti dans G_l. Reprenant (1), d'après le lemme, on obtient l'inégalité

$$\operatorname{Var} S_N(f) \geqslant \sum_{l=1}^{m} \frac{1}{N_l} E_l(\mu_l f(P) - E_l(\mu_l f(P)))^2,$$

or,

$$E_l(\mu_l f(P)) = I_l(f), \quad E_l(\mu_l f(P) - I_l(f))^2 = D_l(f),$$

et le deuxième membre de la dernière inégalité coïncide avec $\operatorname{Var}(\overline{S}_N(f))$. Donc, la relation demandée $\operatorname{Var}(S_N(f)) \geqslant \operatorname{Var}(\overline{S}_N(f))$ est démontrée.

Le fractionnement du domaine d'intégration pour diminuer la variance de la méthode de Monte-Carlo est largement utilisé dans le traitement de l'information scientifique. Soit à déterminer les réserves de neige dans le bassin d'un fleuve. En appliquant directement la méthode de Monte-Carlo on choisirait de façon indépendante plusieurs points à densité de répartition uniforme en lesquels on mesurerait la quantité de neige par unité de superficie. Pour les parcelles de surface aux conditions naturelles identiques (altitude, niveau de nébulosité, peuplement forestier, direction des versants, précipitations, direction générale du vent) cette quantité de neige est approximativement égale. C'est pourquoi la division du bassin en de telles parcelles permet d'améliorer sensiblement la précision.

§ 12. Formules de quadrature de précision élevée aux points aléatoires

Si la fonction sous le signe somme est régulière, la division du domaine d'intégration en parties augmente l'ordre de grandeur de la vitesse de convergence. Soit à calculer

$$I(f) = \int_0^1 \dots \int_0^1 f(x_1, \dots, x_s)\, dx_1 \dots dx_s.$$

Posons $N = n^s$ et partageons le domaine initial d'intégration en cubes égaux Π_m:

$$\frac{m_1 - 1}{n} \leqslant x_1 \leqslant \frac{m_1}{n}, \dots, \frac{m_s - 1}{n} \leqslant x_s \leqslant \frac{m_s}{n}.$$

Dans chaque cube choisissons un point aléatoire P_m. Prenons sa densité de répartition pour une constante égale à n^s et supposons que les points sont deux à deux indépendants. Posons

$$\overline{S}_N(f) = \frac{1}{n^s} \sum_{m_1, \dots, m_s = 1}^{n} f(P_m).$$

La condition $N_l = N\mu_l\mu^{-1}$ est remplie dans ce cas. Par suite, d'après la relation démontrée plus haut, on a

$$\operatorname{Var}\overline{S}_N(f) \leqslant \operatorname{Var} S_N(f) = \frac{D(f)}{N},$$

de plus,

$$\mathrm{E}\overline{S}_N(f) = I(f).$$

Précisons l'estimation de la variance en supposant que la fonction $f(P)$ satisfait à la condition de Lipschitz dans le rapport A pour chaque argument. L'égalité

$$\operatorname{Var}\overline{S}_N(f) = \sum_{m_1, \ldots, m_s=1}^{n} \operatorname{Var}\left(\frac{1}{n^s} f(P_{\mathbf{m}})\right) =$$

$$= \sum_{m_1, \ldots, m_s=1}^{n} \frac{1}{n^{2s}} \operatorname{Var} f(P_{\mathbf{m}}) \qquad (1)$$

est vérifiée.

Evaluons de façon grossière $\operatorname{Var} f(P_{\mathbf{m}})$; il vient

$$\mathrm{E}f(P_{\mathbf{m}}) = \sigma_{\mathbf{m}} = \int_{\Pi_{\mathbf{m}}} n^s f(P)\, dP ;$$

par suite du théorème de la moyenne

$$\sigma_{\mathbf{m}} = f(\overline{P}_{\mathbf{m}}), \quad \text{où} \quad \overline{P}_{\mathbf{m}} \in \Pi_{\mathbf{m}}$$

et donc

$$\mathrm{E}f(P_{\mathbf{m}}) = f(\overline{P}_{\mathbf{m}}).$$

Par ailleurs,

$$\begin{aligned} f(x_1 + \Delta_1, \ldots, x_s + \Delta_s) - f(x_1, \ldots, x_s) = \\ = (f(x_1 + \Delta_1, \ldots, x_s + \Delta_s) - f(x_1 + \Delta_1, \ldots, x_{s-1} + \\ + \Delta_{s-1}, x)) + (f(x_1 + \Delta_1, \ldots, x_{s-1} + \Delta_{s-1}, x_s) - \\ - f(x_1 + \Delta_1, \ldots, x_{s-2} + \Delta_{s-2}, x_{s-1}, x_s)) + \ldots \\ \ldots + (f(x_1 + \Delta_1, x_2, \ldots, x_s) - f(x_1, \ldots, x_s)). \end{aligned}$$

Il en découle que pour les fonctions de la classe en question on a

$$|f(x_1 + \Delta_1, \ldots, x_s + \Delta_s) - f(x_1, \ldots, x_s)| \leqslant$$
$$\leqslant A(|\Delta_1| + \ldots + |\Delta_s|).$$

Si le point $P_{\mathbf{m}}$ appartient à $\Pi_{\mathbf{m}}$, alors chacune de ses coordonnées diffère de la coordonnée correspondante du point $\overline{P}_{\mathbf{m}}$ de n^{-1} au plus. C'est pourquoi la dernière inégalité entraîne l'estimation

$$|f(P_{\mathbf{m}}) - f(\overline{P}_{\mathbf{m}})| \leqslant Asn^{-1}$$

pour $P_m \in \Pi_m$, d'où

$$|f(P_m) - \sigma_m| \leqslant An^{-1}. \tag{2}$$

Toute variable aléatoire ξ satisfait à l'inégalité

$$|E\xi| \leqslant \sup |\xi|.$$

Donc

$$\operatorname{Var}(f(P_m)) = E(f(P_m) - \sigma_m)^2 \leqslant \sup_{P_m \in \Pi_m} (f(P_m) - \sigma_m)^2 \leqslant \left(\frac{As}{n}\right)^2.$$

A l'aide de cette estimation on conclut que le deuxième membre de (1) ne dépasse pas la quantité

$$n^s n^{-2s} \left(\frac{As}{n}\right)^2 = \frac{(As)^2}{n^{s+2}}.$$

Ayant désigné par N le nombre total des points n^s, on obtient l'estimation

$$\operatorname{Var} \bar{S}_N(f) \leqslant \frac{A^2 s^2}{N^{1+2/s}}. \tag{3}$$

D'où, d'après l'inégalité de Tchebycheff (9.1), il vient

$$|\bar{S}_N(f) - I(f)| \leqslant \frac{As}{N^{1/s}\sqrt{\eta N}}$$

avec la probabilité $1 - \eta$. Du point de vue de l'ordre de grandeur l'estimation obtenue est meilleure que celle $O(1/\sqrt{N})$ de la méthode de Monte-Carlo. Il s'agit de l'estimation en probabilité ; il se trouve que pour la méthode considérée on peut obtenir une estimation garantie. La multiplication de (2) par n^{-s} donne

$$\left|\frac{1}{n^s} f(P_m) - \int_{\Pi_m} f(P)\, dP\right| \leqslant \frac{As}{n^{s+1}}.$$

La quantité $\bar{S}_N(f) - I(f)$ peut être représentée sous la forme d'une somme des termes

$$\sum_{m_1, \ldots, m_s = 1}^{n} \left(\frac{1}{n^s} f(P_m) - \int_{\Pi_m} f(P)\, dP\right).$$

En sommant les estimations pour ces termes, on obtient

$$|\bar{S}_N(f) - I(f)| \leqslant \frac{As}{n} = \frac{As}{N^{1/s}}.$$

La comparaison avec le théorème du § 7 montre que la méthode en question admet une estimation garantie de l'erreur, qui est optimale relativement à l'ordre sur la classe considérée.

Peut-on améliorer l'estimation de variance (3) sur cette classe? La méthode de Monte-Carlo et les autres procédés d'intégration analogues au mode opératoire considéré et tels que la valeur approchée de l'intégrale dépende de certains paramètres aléatoires sont dites *indéterministes*. Soit $S_N(F)$ la valeur approchée de l'intégrale $I(f)$ obtenue par une méthode indéterministe.

THÉORÈME (sans démonstration). *Il existe* $d_1(\mathrm{r}, \mathrm{A})$, $d_2 > 0$, *satisfaisant à la relation suivante. Pour toute méthode de calcul d'intégrales qui n'utilise de l'information sur la fonction sous le signe somme que celle sur ses valeurs en N points, il existe une fonction* $f \in C_{\mathrm{r}}(\mathrm{A})$ *telle que*

$$|S_N(f) - I(f)| > \frac{d_1(\mathrm{r}, \mathrm{A})}{N^{r+1/2}} \tag{4}$$

avec la probabilité d_2; *ici comme plus haut* $\frac{1}{r} = \frac{1}{r_1} + \ldots + \frac{1}{r_s}$, $S_N(f)$ *est une valeur approchée de l'intégrale.*

Il est évident que (4) entraîne l'inégalité

$$\operatorname{Var} S_N(f) \geqslant \frac{d_3(\mathrm{r}, \mathrm{A})}{N^{2r+1}}$$

qui signifie en particulier que l'ordre de grandeur de (3) ne peut être amélioré.

THÉORÈME (sans démonstration). *On peut indiquer des procédés d'intégration pour lesquels on a simultanément l'estimation d'erreur garantie*

$$|S_N(f) - I(f)| \leqslant \frac{d_4(\mathrm{r}, \mathrm{A})}{N^r}$$

et

$$58 \quad \mathrm{E}S_N(f) = I(f), \ \sqrt{\operatorname{Var} S_N(f)} \leqslant \frac{d_5(\mathrm{r}, \mathrm{A})}{N^{r+1/2}}$$

quelles que soient $f \in C_{\mathrm{r}}(\mathrm{A})$.

Nous avons construit plus haut un tel procédé pour $r_1 = \ldots = r_s = 1$. Il consiste à partager le domaine initial d'intégration en plusieurs parties et à calculer l'intégrale sur chacune d'elles au moyen d'une « formule de quadrature aléatoire ».

Considérons la classe de fonctions $C_{(2, \ldots, 2)}(A, \ldots, A)$. Comme au début de ce paragraphe prenons dans chaque cube Π_{m} un point aléatoire P_{m}. Désignons par P_{m}^* le point symétrique de P_{m} par rapport au centre du cube; posons

$$s_{\mathrm{m}}(f) = \frac{1}{2n^s}(f(P_{\mathrm{m}}) + f(P_{\mathrm{m}}^*)),$$

$$S_N(f) = \sum_{m_1, \ldots, m_s = 1}^{n} s_{\mathrm{m}}(f).$$

Les points P_m, P_m^* sont uniformément répartis dans le cube considéré; donc,

$$\mathrm{E}\left(\frac{1}{n^s} f(P_m)\right) = \mathrm{E}\left(\frac{1}{n^s} f(P_m^*)\right) = \int_{\Pi_m} f(P)\,dP.$$

Il s'ensuit que

$$\mathrm{E}s_m(f) = \int_{\Pi_m} f(P)\,dP$$

la sommation de ces relations nous donne

$$\mathrm{E}S_N(f) = I(f).$$

Comme plus haut, on a

$$\mathrm{Var}\, S_N(f) = \sum_{m_1, \ldots, m_s=1}^{n} \mathrm{Var}\, s_m(f) = \frac{1}{n^{2s}} \sum_{m_1, \ldots, m_s=1}^{n} \mathrm{Var}\frac{(f(P_m)+f(P_m^*))}{2}. \tag{5}$$

Désignons par $\overline{P}_m$ le centre du cube Π_m; en soustrayant la constante sous le signe de variance, on obtient l'égalité

$$\mathrm{Var}\,\frac{(f(P_m)+f(P_m^*))}{2} = \mathrm{Var}\,\frac{(f(P_m)-2f(\overline{P}_m)+f(P_m^*))}{2}. \tag{6}$$

Pour un point $P_m=(x_1, \ldots, x_s)$ fixe, introduisons la fonction

$$g(t) = f\left(\frac{m_1-1/2}{n} + \frac{x_1-(m_1-1/2)}{n}\,t, \ldots, \frac{m_s-1/2}{n} + \frac{x_s-(m_s-1/2)}{n}\,t\right).$$

Il est clair que

$$g(1) = f(P_m), \quad g(0) = f(\overline{P}_m), \quad g(-1) = f(P_m^*).$$

Exprimons la différence seconde en fonction de la dérivée deuxième

$$f(P_m)-2f(\overline{P}_m)+f(P_m^*) = g(1)-2g(0)+g(-1) = g''(\theta), \quad |\theta|<1.$$

Une dérivation immédiate nous montre que

$$g''(t) = \sum_{i,j=1}^{s} f_{x_i x_j}\left(\frac{m_1-1/2}{n} + \frac{nx_1-(m_1-1/2)}{n}\,t, \ldots\right) \times \times \left(x_i - \frac{m_i-1/2}{n}\right)\left(x_j - \frac{m_j-1/2}{n}\right).$$

Puisque $(x_1, \ldots, x_s) \in \Pi_m$,

$$\left|x_i - \frac{m_i-1/2}{n}\right| \leqslant \frac{1}{2n}.$$

C'est pourquoi

$$|g''(t)| \leqslant \frac{As^2}{4n^2} \quad \text{pour } |t| \leqslant 1,$$

$$\frac{|f(P_m)-2f(\overline{P}_m)+f(P_m^*)|}{2} \leqslant \frac{As^2}{8n^2}. \tag{7}$$

Conformément à la dernière inégalité on a

$$\operatorname{Var} s_m(f) \leqslant \frac{1}{n^{2s}} \left(\frac{As^2}{8n^2}\right)^2.$$

La substitution de cette estimation dans le deuxième membre de (5) donne

$$\operatorname{Var} S_N(f) \leqslant \frac{(As^2)^2 2^{4/s}}{32N^{1+4/s}}, \quad N = 2n^s.$$

D'où, à l'aide de l'inégalité de Tchebycheff (9.1), on obtient

$$|S_N(f) - I(f)| \leqslant \frac{As^2 2^{1/2+2/s}}{8N^{2/s}\sqrt{\eta N}}$$

avec la probabilité $1 - \eta$.

Le procédé d'intégration considéré permet également d'obtenir une estimation d'erreur garantie $O(AN^{-2/s})$. En effet, la relation (7) signifie que pour tous les P_m la valeur de $s_m(f)$ appartient au segment

$$\left[\frac{1}{n^s}\left(f(\overline{P}_m) - \frac{As^2}{8n^2}\right), \ \frac{1}{n^s}\left(f(\overline{P}_m) + \frac{As^2}{8n^2}\right)\right].$$

Donc, il en est également de son espérance mathématique

$$\int_{\Pi_m} f(P)\, dP.$$

On en déduit que pour tous les P_m la quantité

$$\left| s_m(f) - \int_{\Pi_m} f(P)\, dP \right|$$

ne dépasse pas la longueur de ce segment, à savoir $\frac{As^2}{4n^{2+s}}$. Ainsi,

$$\left| s_m(f) - \int_{\Pi_m} f(P)\, dP \right| \leqslant \frac{As^2}{4n^{2+s}}.$$

En sommant par rapport à m, on obtient l'estimation d'erreur

$$|S_N(f) - I(f)| \leqslant \frac{As^2}{4n^2} = \frac{As^2 2^{2/s}}{4N^{2/s}}.$$

La comparaison avec les énoncés des théorèmes sur les minorations des estimations garanties et indéterministes d'erreurs montre que dans le cas de la classe $C_{(s,\ldots,s)}(A, \ldots, A)$ et les estimations garanties et celles en probabilité de la vitesse de convergence du procédé construit ci-dessus sont optimales relativement à l'ordre.

§ 13. Choix de la méthode

Discutons le schéma de résolution d'un problème qui s'est très bien prêté aux techniques de Monte-Carlo de ce type. On a eu à calculer la série d'intégrales

$$I(\alpha_1, \alpha_2, \alpha_3) = \int_G f(P; \alpha_1, \alpha_2, \alpha_3)\, dP$$

pour différentes valeurs des paramètres α_1, α_2, α_3. Le domaine d'intégration G était compris dans le cube unité à trois dimensions, la condition d'appartenance d'un point à G étant donnée par un système d'inégalités encombrantes. Il était clair qu'on ne pouvait faire le changement de variables indépendantes de façon que G prît une forme standard permettant l'application de formules de quadrature très précises. On a dû prolonger f par 0 en dehors de G et ramener le problème initial au calcul de l'intégrale

$$\overline{I}(\alpha_1, \alpha_2, \alpha_3) = \int_0^1 \int_0^1 \int_0^1 \overline{f}(P\,;\, \alpha_1, \alpha_2, \alpha_3)\, dx_1\, dx_2\, dx_3,$$

où

$$\overline{f}(P\,;\, \alpha_1, \alpha_2, \alpha_3) = \begin{cases} f(P\,;\, \alpha_1, \alpha_2, \alpha_3) & \text{pour } P \in G, \\ 0 & \text{pour } P \notin G. \end{cases}$$

La fonction à intégrer était désormais discontinue. On a calculé l'intégrale par la formule des rectangles et celle de Gauss. Pour vérifier la précision on a fait les calculs avec un nombre variable de points d'intégration ; les résultats se sont stabilisés lentement (ils variaient rapidement avec chaque nouveau nombre de points), ce qui indiquait une précision faible des valeurs approchées obtenues. L'application immédiate de la méthode de Monte-Carlo a entraîné une stabilisation encore plus fâcheuse des valeurs approchées. On a donc décidé de faire recours aux procédés décrits de réduction de la variance par division du domaine d'intégration. On a partagé le domaine $0 \leqslant x_1, x_2, x_3 \leqslant 1$ en cubes égaux de côté $1/n$, puis on a essayé les deux méthodes considérées au § 12. Comme dans le cas d'autres formules de quadrature, c'est la stabilisation des résultats qui nous intéressait. Il s'est trouvé que les deux méthodes pouvaient assurer la précision demandée des calculs en temps machine admissible.

Dans de nombreux cas le calcul d'intégrales de multiplicité élevée ou d'une grande série d'intégrales ne peut être effectué que grâce à l'élimination des calculs analogues répétitifs. La résolution du problème considéré a d'abord paru inextricable, tellement le test d'appartenance du point P au domaine G était laborieux ; la recherche de chaque valeur $f(P\,;\, \alpha_1, \alpha_2, \alpha_3)$ exigeait notablement moins de calculs. Puisque toutes les intégrales étaient calculées sur un même domaine, on a décidé d'employer l'algorithme suivant. On a partagé le cube $0 \leqslant x_1, x_2, x_3 \leqslant 1$ en petits cubes de côté $1/n$, dans chacun desquels on a choisi un point aléatoire P_m et vérifié la condition $P_m \in G$ (les coordonnées de tous les points $P_m \in G$ étaient stockées en mémoire de l'ordinateur). Ensuite on a remplacé toutes les intégrales de la série par les sommes

$$\sum_{P_m \in G} \frac{1}{n^3} f(P_m\,;\, \alpha_1, \alpha_2, \alpha_3)$$

ayant les mêmes points. Comme on n'avait plus à tester la condition $P_m \in G$ lors du calcul de chaque intégrale, on a réduit d'environ 100 fois le temps machine. Et ce n'était pas tout. La fonction sous le signe somme étant

$$g(x_1, \alpha_1)\, h(x_1, x_2, x_3\,;\, \alpha_1, \alpha_2, \alpha_3),$$

où le calcul de chaque valeur de la fonction h exigeait relativement peu d'opérations élémentaires sur ordinateur devant celui de g, on a employé le schéma suivant. Toutes les intégrales $\overline{I}(\alpha_1, \alpha_2, \alpha_3)$ correspondant à une même valeur du paramètre α_1 ont été obtenues simultanément, et, en travaillant sur toute la série, chaque valeur $g(x_1, \alpha_1)$ a donc été trouvée une seule fois. Cet arrangement des calculs a permis d'abaisser le temps machine à un niveau admissible. Une fois la série calculée, on a analysé la marche des opérations et découvert certaines possibilités de faciliter la tâche du client et de l'utilisateur. Pour

réduire le temps machine le client voulait diminuer le nombre M des triplets de valeurs $(\alpha_1^i, \alpha_2^i, \alpha_3^i)$ pour lesquels on cherchait la valeur de l'intégrale. L'opération la plus laborieuse était la recherche des valeurs de la fonction $g(x_1, \alpha_1)$. C'est pourquoi le coût total en temps machine était proportionnel au nombre des diverses valeurs α_1^i et non pas à M. Ainsi, on aurait pu le réduire également par un choix heureux du triplet $(\alpha_1^i, \alpha_2^i, \alpha_3^i)$.

Le problème en question offrait encore un moyen d'élever la précision. L'effort de calcul qu'exige la méthode considérée est proportionnel au produit du nombre des α différents par celui des différentes valeurs des x_1, coordonnées des points d'intégration. Ceux-ci étant choisi au hasard, ce dernier nombre était très élevé. Pour le diminuer on aurait pu procéder de la façon suivante: approcher l'intégrale initiale par la formule des rectangles

$$\overline{I}(\alpha_1, \alpha_2, \alpha_3) \approx \sum_{q=1}^{m} \frac{1}{m} I\left(\alpha_1, \alpha_2, \alpha_3, \frac{q-1/2}{m}\right),$$

où

$$I\left(\alpha_1, \alpha_1, \alpha_3, \frac{q-1/2}{m}\right) = \int_0^1 \int_0^1 \overline{f}\left(\alpha_1, \alpha_2, \alpha_3, \frac{q-1/2}{m}, x_2, x_3\right) dx_2\, dx_3,$$

et n'employer la méthode de Monte-Carlo avec division du domaine d'intégration que pour trouver $I(\alpha_1, \alpha_2, \alpha_3; x_1)$. Un autre procédé consiste à diviser le segment $[0, 1]$ en segments partiels $\left[\frac{q-1}{m}, \frac{q}{m}\right]$, à choisir sur chacun d'eux un point aléatoire ξ_q et à poser

$$\overline{I}(\alpha_1, \alpha_2, \alpha_3) \approx \frac{1}{m} \sum_{q=1}^{m} I(\alpha_1, \alpha_2, \alpha_3; \xi_q).$$

Ces deux méthodes convergent plus lentement, mais elles exigent sensiblement moins de valeurs $g(x_1, \alpha_1)$.

On voit donc qu'en calculant les grandes séries d'intégrales (comme, d'ailleurs, les grandes séries de n'importe quels problèmes), ce n'est pas souvent l'amélioration de la méthode dans chaque problème de la série qui donne de bons résultats, mais une meilleure organisation des calculs.

En utilisant le procédé décrit au § 5 le calcul d'une intégrale de multiplicité élevée se ramène en quelque sorte à celui d'une série d'intégrales plus simples. C'est pourquoi dans le cas d'intégrales de haute multiplicité on augmentera l'efficacité par les techniques décrites ci-dessus.

A noter le risque suivant si l'on veut résoudre toute la série des problèmes à la fois.

L'ensemble des problèmes proposés correspond à un phénomène réel. Il se peut que ce phénomène n'ait pas encore été calculé sur ordinateur et que le modèle mathématique ne soit pas satisfaisant. Si l'on résout alors simultanément tous les problèmes de la série,

les résultats s'avéreront inutiles. Tandis qu'une résolution de proche en proche révèle dès les premiers résultats un écart entre le modèle mathématique et la nature générale du phénomène. Un tel écart est inhérent à la résolution de nouveaux problèmes et il ne s'élimine qu'après un grand nombre de calculs d'essai et une étude commune du modèle par l'utilisateur et le client.

L'utilisateur a toutes les raisons de vouloir participer à la discussion de la position du problème.

En effet, si la position initiale du problème est déraisonnable, il perdra beaucoup de temps à choisir un nouvel algorithme et à écrire un nouveau programme. Tandis qu'en prenant part à la discussion, il établit pour lui les variantes de la position, dont il tiendra compte en rédigeant le programme. Il importe donc beaucoup que dans le cas de problèmes d'un type nouveau le programme se compose de divers blocs fonctionnels susceptibles d'une modification indépendante.

Le conflit d'intérêts du client (calcul d'un modèle détaillé) et de l'utilisateur (volume minimal de calculs) engendre souvent un modèle simplifié du phénomène dont le calcul rapide permet de décider si l'étude d'un modèle plus compliqué est utile.

Considérons un autre exemple relatif au choix de la méthode d'intégration dans le cas d'intégrales multiples. On ramène habituellement le calcul des intégrales sur un domaine compliqué à celui des intégrales prises sur un produit direct de domaines simples: segments, parallélépipèdes, rayons, droites infinies, cercles, sphères, boules, etc.

Il y a assez de méthodes parfaites d'intégration pour ces domaines standard (p. ex. le procédé décrit au § 5 ou un autre analogue).

Soit à calculer l'intégrale

$$I(f) \equiv \iiint\limits_{1 \leqslant x_1^2 + x_2^2 + x_3^2 \leqslant 4} f(x_1, x_2, x_3)\, dx_1\, dx_2\, dx_3.$$

Il est commode de l'écrire sous la forme

$$I(f) = \int\limits_1^2 \int\limits_{S_1} g(l, \omega)\, d\omega\, dl,$$

S_1 étant la sphère unité et $d\omega$ l'élément d'aire,

$$g(l, \omega) = l^2 f(l\omega_1, l\omega_2, l\omega_3),$$
$$\omega = (\omega_1, \omega_2, \omega_3), \quad \omega_1^2 + \omega_2^2 + \omega_3^2 = 1.$$

Supposons qu'on désire calculer cette intégrale au moyen des formules de quadrature « produit direct » des formules sur le segment [1, 2] et la sphère S_1.

On comprend par *produit direct* de la formule de quadrature

$$\int_1^2 h(l)\,dl \approx \sum_{j=1}^{n} d_j h(l_j) \tag{1}$$

et de

$$\int_{S_1} p(\omega)\,d\omega \approx \sum_{q=1}^{m} k_q p(\omega_q) \tag{2}$$

la formule

$$\int_1^2 \int_{S_1} g(l,\ \omega)\,d\omega\,dl \approx \sum_{j=1}^{n} \sum_{q=1}^{m} d_j k_q g(l_j,\ \omega_q).$$

Etablissons le lien existant entre l'erreur d'intégration, d'une part, et la technique d'intégration et le nombre des points, de l'autre. Etudions le comportement de l'erreur d'intégration numérique sur l'axe l à l'aide d'une formule de quadrature « de base » sur la sphère unité

$$\int_{S_1} p(\omega)\,d\omega \approx \sum_{q=1}^{m_0} k_q^0 p(\omega_q^0); \tag{3}$$

on exige que le nombre m_0 des points soit petit et qu'en même temps les expressions

$$G(l) = \int_{S_1} g(l,\ \omega)\,d\omega \quad \text{et} \quad G^0(l) = \sum_{q=1}^{m_0} k_q^0 g(l,\ \omega_q^0)$$

présentent sur l une allure qualitative identique. Par exemple, on pourrait essayer de prendre pour (3) la formule

$$\int_{S_1} p(\omega)\,d\omega \approx \frac{2\pi}{3} \sum p(\omega_1.\ \omega_2,\ \omega_3),$$

où la sommation se fait sur les 6 points d'intersection de la sphère unité avec les axes de coordonnées x_1, x_2, x_3.

Supposons que pour calculer l'intégrale sur l'axe on emploie la formule de Gauss ou celle de Simpson. En utilisant successivement les formules de Gauss

$$\int_1^2 h(l)\,dl \approx \sum_{j=1}^{n} d_j^n h(l_j^n)$$

pour $n = n_1, n_2, \ldots$ points, on obtient des quantités

$$G^0_{n_l} = \sum_{j=1}^{n_l} d_j^{n_l} G^0 (l_j^{n_l}).$$

De même, pour $n = n'_1, n'_2, \ldots$ points on trouve des approximations à l'aide de la formule de Simpson $S^0_{n'_i}$.

L'étude du comportement de toutes ces quantités donne la valeur de leur limite I^0. Choisissons ensuite, pour chaque n, parmi toutes les approximations $G^0_{n_l}$ et $S^0_{n'_i}$ avec $n_i, n'_i \leqslant n$ l'approximation Γ_n assurant la meilleure précision et introduisons la fonction de l'erreur d'intégration numérique sur l'axe l

$$\varphi_l (n) = |\Gamma_n - I^0|.$$

De même, on fixe une formule de quadrature de base par rapport à l'argument l (c'est souvent la formule de Gauss à deux points) et on construit la fonction $\varphi_\omega (m)$ de l'erreur résultant de l'intégration numérique sur la sphère ω. Supposons que l'erreur résultante est $R = \varphi_l (n) + \varphi_\omega (m)$. Si les valeurs de la fonction f sont calculées de façon indépendante, le coût de la méthode est proportionnel à nm. En minimisant nm dans la condition $\varphi_l (n) + \varphi_\omega (m) \leqslant \varepsilon$, on obtient les valeurs n_0 et m_0 du nombre de points cherchés. On choisit la formule (celle de Gauss ou de Simpson) à laquelle correspond n donné. En cas de doute sur la légitimité de cette technique il est possible de vérifier l'exactitude du résultat par une intégration complémentaire avec des valeurs de n et m légèrement différentes de n_0 et m_0.

Voici un exemple typique d'un tel choix de la méthode d'intégration pour chaque axe dans le cas d'une grande série d'intégrales prises sur le cube unité de dimension s. On intègre sur chaque axe par la formule de Gauss; en choisissant le nombre de ses points sur chaque axe on prend pour la formule de base, sur les axes restants de la formule de Gauss à deux points. Le fractionnement du nombre de points sur chaque axe considéré se poursuit jusqu'à ce que la différence entre deux approximations consécutives ne soit inférieure à $\varepsilon \cdot s^{-1}$. Après avoir déterminé le nombre nécessaire $n^{(i)}$ de points correspondant à chaque axe, on calcule l'intégrale avec $n^{(i)}$ points sur chaque axe.

L'algorithme décrit n'étant pas tout à fait sûr, on le teste habituellement *pour une série donnée d'intégrales*: on choisit une sous-série assez représentative et on compare les résultats des calculs selon l'algorithme donné avec ceux pour des valeurs légèrement différentes du nombre de points sur les axes.

Dans un des programmes standard de la méthode de Gauss à plusieurs dimensions le nombre $n^{(i)}$ de points sur chaque axe est choisi selon le procédé décrit ci-dessus, puis on exécute un calcul correspondant à ces $n^{(i)}$ et un calcul de vérification pour des valeurs de $n^{(i)}$ légèrement différentes.

Bibliographie du chapitre V

I [6], II [13, 14]; §§ 1-4: II [16, 37, 39], III [24, 25, 33, 34, 35, 44, 45]; § 4: III [4]; § 7: III [10]; § 8: III [13]; §§ 9-12: II [2, 8, 18], III [10, 61]; § 13: II [9].

DEUXIÈME PARTIE

PROBLÈMES D'ALGÈBRE ET D'OPTIMISATION

CHAPITRE VI

MÉTHODES NUMÉRIQUES DE L'ALGÈBRE

L'objet des méthodes numériques de l'algèbre est la résolution de systèmes d'équations linéaires, l'inversion des matrices, le calcul des déterminants, la recherche des vecteurs et valeurs propres de matrices et des zéros de polynômes.

Abordés formellement, ces problèmes se résolvent sans difficulté : on résout un système en développant les déterminants dans la formule de Cramer ; on trouve les valeurs propres d'une matrice en écrivant l'équation caractéristique et en en obtenant les racines. Cette façon d'agir appelle cependant des objections.

Ainsi, la résolution d'un système à m inconnues par développement direct des déterminants exige des opérations arithmétiques de l'ordre de $m!m$, ce qui dépasse la capacité des ordinateurs modernes dès que $m = 30$. L'exemple suivant prouve que pour un m suffisamment grand il en sera ainsi dans l'avenir immédiat.

Imaginons-nous une machine parallèle occupant un volume V et composée d'éléments dont chacun est un cube de côtés égaux Δ. On admet naturellement que le temps mis par un élément quelconque pour effectuer une opération élémentaire est au moins Δ/c, $c = 3 \cdot 10^8$ m/s étant la vitesse de propagation de la lumière, i.e. la plus grande vitesse possible de transmission de l'information. On majore alors par Vc/Δ^4 le nombre total d'opérations élémentaires exécutables en une seconde par la machine donnée. Si $V = 1$ km^3 et $\Delta = 10^{-8}$ cm (ordre de grandeur du rayon atomique), ce nombre vaut $3 \cdot 10^{57}$. Or, $100! = 10^{157,9}\ldots$

On voit donc qu'il sera impossible de calculer explicitement un déterminant d'ordre 100.

Mais ces méthodes classiques sont à rejeter même pour un m faible, et la cause en est la grande influence des arrondis sur le résultat final. Dès que $m = 20$, le dépassement de capacité des nombres détermine l'arrêt de l'ordinateur, et même si l'arrêt n'a pas lieu, l'erreur de calcul fait que le résultat présente souvent un écart sensible à la valeur exacte. Il en est de même pour la recherche des valeurs propres des matrices en explicitant le polynôme caractéristique.

Les méthodes de résolution des problèmes algébriques peuvent être directes, itératives ou probabilistes. Les classes de problèmes qui sont résolus par ces méthodes sont conventionnellement appelées classes de problèmes à petit, moyen ou grand nombre d'inconnues respectivement.

La modification du volume et de la structure de la mémoire des calculateurs, l'augmentation de leur rapidité de fonctionnement et le développement des procédés de calcul entraînent un décalage des domaines d'applications de ces méthodes en direction des systèmes d'ordre plus grand. Les méthodes directes sont valables actuellement pour des systèmes de l'ordre de 10^3 et les techniques itératives pour ceux de l'ordre de 10^6.

§ 1. Méthodes d'élimination successive des inconnues

Considérons le problème de résolution du système $A\mathbf{x} = \mathbf{b}$; ici $A = [a_{ij}]$ est une matrice $m \times m$, $\det A \neq 0$

$$b = (a_{1,\,m+1},\ \ldots,\ a_{m,\,m+1})^T.$$

Une méthode de résolution est directe si elle donne, au bout d'un nombre fini d'opérations logiques et arithmétiques et dans l'hypothèse de l'absence d'arrondis, une solution exacte du problème. Si le nombre d'éléments non nuls de la matrice du système est de l'ordre de m^2, le calcul par les méthodes directes actuellement disponibles exige un nombre d'opérations de l'ordre de m^3. Donc, pour que les méthodes directes passent, il est nécessaire que cet ordre de grandeur soit acceptable pour la machine donnée; il existe également des contraintes qu'imposent le volume et la structure de la mémoire du calculateur.

On connaît des méthodes de résolution de pareils systèmes qui sont moins coûteuses en opérations; cependant, le résultat qu'elles permettent d'obtenir est très sensible à l'erreur de calcul, ce qui fait qu'elles ne sont pas exploitables.

Parmi les procédés directs la *méthode d'élimination de Gauss* est la plus connue. Voyons l'une de ses réalisations possibles. Dans l'hypothèse $a_{11} \neq 0$, divisons la première équation du système

$$\sum_{j=1}^{m} a_{ij}x_j = a_{i,\,m+1}, \quad i = 1,\ \ldots,\ m, \tag{1}$$

par le coefficient a_{11}, d'où

$$x_1 + \sum_{j=2}^{m} a_{1j}^1 x_j = a_{1,\,m+1}^1.$$

Retranchons ensuite de chacune des équations restantes la première multipliée par le coefficient a_{i1} correspondant. On a

$$\sum_{j=2}^{m} a_{ij}^{1} x_j = a_{i,\,m+1}^{1}, \quad i = 2, \ldots, m.$$

La première inconnue s'est éliminée entre toutes les équations, excepté la première. Supposons ensuite que $a_{22}^{1} \neq 0$, divisons la deuxième équation par a_{22}^{1} et éliminons l'inconnue x_2 entre toutes les équations à partir de la deuxième, et ainsi de suite. L'élimination successive des inconnues met le système sous forme triangulaire

$$x_i + \sum_{j=i+1}^{m} a_{ij}^{i} x_j = a_{i,\,m+1}^{i}, \quad i = 1, \ldots, m. \tag{2}$$

x_i s'élimine à la suite des opérations suivantes: 1) division de la i-ième équation par a_{ii}^{i-1}; 2) soustraction de la i-ième équation multipliée par a_{ik}^{i-1} des équations de numéros $k = i + 1, \ldots, m$. La première opération équivaut à multiplier à gauche le système par la matrice diagonale

$$C_i = \begin{pmatrix} 1 & & & & & & \\ & \ddots & & & & & \\ & & 1 & & & & \\ & & & (a_{ii}^{i-1})^{-1} & & & \\ & & & & 1 & & \\ & & & & & \ddots & \\ & & & & & & 1 \end{pmatrix};$$

la seconde à multiplier à gauche par 58

$$\widetilde{C}_i = \begin{pmatrix} 1 & & & & & & \\ & \ddots & & & & & \\ & & 1 & & & & \\ & & \alpha_{i+1} & 1 & & & \\ & & \vdots & & \ddots & & \\ & & \alpha_m & & & 1 \end{pmatrix},$$

où $\alpha_k = -a_{i,\,k}^{i-1}$, $i < k \leqslant m$.

Le système (2), résultat de ces transformations, s'écrit donc

$$CA\mathbf{x} = C\mathbf{b}$$

avec

$$C = C_m\widetilde{C}_{m-1}C_{m-1} \ldots \widetilde{C}_1C_1.$$

Le produit de matrices triangulaires gauches (resp. droites) est encore une matrice triangulaire gauche (resp. droite); C est donc triangulaire gauche.

La formule des éléments d'une matrice inverse

$$(A^{-1})_{ij} = \frac{A_{ji}}{\det A}$$

entraîne que l'inverse d'une matrice triangulaire gauche (resp. droite) est triangulaire gauche (resp. droite). Par conséquent, $B = C^{-1}$ est triangulaire gauche.

Introduisons la notation $CA = D$; par construction tous les $d_{ii} = 1$ et D est triangulaire droite. On a

$$A = C^{-1}D = BD.$$

L'égalité $A = BD$ plus les conditions $d_{ii} = 1$, $i = 1, \ldots, m$, forment un système d'équations relativement aux éléments des matrices triangulaires B et D: $\sum_{j=1}^{m} b_{ij}d_{jk} = a_{ik}$. Vu que $b_{ij} = 0$ pour $i < j$ et $d_{jk} = 0$ pour $k < j$, ce système peut prendre la forme

$$\sum_{j=1}^{\min(i,\,k)} b_{ij}d_{jk} = a_{ik} \tag{3}$$

ou, ce qui revient au même,

$$\sum_{j=1}^{k} b_{ij}d_{jk} = a_{ik} \text{ pour } k \leqslant i,$$

$$\sum_{j=1}^{i} b_{ij}d_{jk} = a_{ik} \text{ pour } i < k;$$

mettant à profit la condition $d_{ii} = 1$, on trouve un système de relations de récurrence pour déterminer b_{ij} et d_{ij}:

$$\begin{aligned} b_{ik} &= a_{ik} - \sum_{j=1}^{k-1} b_{ij}d_{jk} \text{ pour } k \leqslant i, \\ d_{ik} &= \frac{a_{ik} - \sum_{j=1}^{i-1} b_{ij}d_{jk}}{b_{ii}} \text{ pour } i < k. \end{aligned} \tag{4}$$

Les calculs s'effectuent successivement pour les ensembles

$$(i, k) = (1, 1), \ldots, (1, m), (2, 1), \ldots, (2, m), \ldots \\ \ldots, (m, 1), \ldots, (m, m).$$

Ici et dans la suite, on estime que la somme totale est nulle si la limite supérieure de sommation est plus petite que la limite inférieure.

Ainsi, au lieu de transformer successivement le système (1) jusqu'à obtenir (2) on peut calculer directement les matrices B et D par les formules (4), à condition toutefois que tous les b_{ii} soient non nuls. Soient A_k, B_k, D_k les matrices des mineurs principaux d'ordre k des matrices A, B, D. Conformément à (3), $A_k = B_k D_k$; puisque $\det D_k = 1$, $\det B_k = b_{11} \ldots b_{kk}$, on a $\det A_k = b_{11} \ldots b_{kk}$. Par conséquent,

$$b_{kk} = \frac{\det A_k}{\det A_{k-1}}, \quad \det A_0 = 1.$$

On voit que pour utiliser les formules (4) il faut et il suffit que

$$\det A_k \neq 0, \quad k = 1, \ldots, m - 1. \tag{5}$$

Ces calculs demandent $(2m^3 - 3m^2 + m)/6$ multiplications, autant d'additions et $m(m - 1)/2$ divisions. Adoptons la notation $C\mathbf{b} = \mathbf{d} = (d_{1,\, m+1}, \ldots, d_{m,\, m+1})^T$. Du moment que $C^{-1} = B$ et $CA = D$, on a $B\mathbf{d} = \mathbf{b}$, $D\mathbf{x} = \mathbf{d}$. Ainsi, après avoir décomposé la matrice du système initial en produit d'une matrice triangulaire gauche par une droite $A = BD$, on ramène la résolution du système initial à la résolution successive de deux systèmes $B\mathbf{d} = \mathbf{b}$ et $D\mathbf{x} = \mathbf{d}$ de matrices triangulaires; le processus coûtera $2m^2 - m$ opérations arithmétiques.

On réunit parfois les opérations successives portant sur la décomposition de A en produit de matrices triangulaires et sur la détermination du vecteur $\mathbf{d}$. Les équations

$$\sum_{j=1}^{i} b_{ij} d_{j,\, m+1} = a_{i,\, m+1}$$

du système $B\mathbf{d} = \mathbf{b}$ se mettent sous la forme (3):

$$\sum_{j=1}^{\min(i,\, m+1)} b_{ij} d_{j,\, m+1} = a_{i,\, m+1}.$$

C'est pourquoi les valeurs $d_{i,m+1}$ se calculent simultanément avec les autres d_{ij} par les formules (4). L'ensemble des opérations transformant le système de départ en $D\mathbf{x} = \mathbf{d}$ constitue le *calcul direct par la méthode de Gauss* et la procédure de résolution de $D\mathbf{x} = \mathbf{d}$ le *calcul par remontée*.

En calcul manuel, la probabilité d'erreurs accidentelles est élevée. Pour éliminer celles-ci on introduit la *colonne de contrôle du système*: $\mathbf{a}_{m+2} = (a_{1,\,m+2}, \ldots$ $\ldots, a_{m,\,m+2})^T$ composée des éléments de contrôle des équations du système

$$a_{l,\,m+2} = \sum_{j=1}^{m+1} a_{lj}.$$

Lors des transformations d'équations, les éléments de contrôle subissent les mêmes opérations que les termes constants. L'élément de contrôle de toute nouvelle équation doit donc être égal à la somme des coefficients de celle-ci; un écart important révélera des erreurs commises au cours des calculs ou l'instabilité de l'algorithme de calcul par rapport à l'erreur de calcul.

Citons un exemple. En réduisant le système d'équations $A\mathbf{x} = \mathbf{b}$ à la forme $D\mathbf{x} = \mathbf{d}$ à l'aide des formules (4), l'élément de contrôle $d_{l,\,m+2}$ de chacune des équations du système $D\mathbf{x} = \mathbf{d}$ s'obtient par les mêmes formules (4). Après avoir calculé tous les d_{ij} (i fixe), on teste l'égalité

$$\sum_{j=i}^{m+1} d_{ij} = d_{i,\,m+2}.$$

Le calcul par remontée comporte lui aussi le calcul des éléments de contrôle des lignes.

Dans plusieurs cas, on sait d'avance que la condition (5) est remplie; de nombreux problèmes de physique mathématique se ramènent par exemple à la résolution de systèmes ayant une matrice A définie positive.

Cela n'est cependant pas vrai dans le cas général. En effet, il se peut que tous les $\det A_{kk} \neq 0$, mais certains b_{kk} sont très petits; une division par ces derniers donnera des nombres importants affectés d'erreurs absolues élevées, et le résultat s'en trouvera fortement altéré.

Pour éviter l'influence catastrophique de l'erreur de calcul, on applique la *méthode de Gauss avec choix du pivot*. Voyons en quoi diffère-t-elle du schéma décrit ci-dessus. Supposons qu'en cours d'élimination des inconnues on obtienne le système

$$x_i + \sum_{j=i+1}^{m} a_{ij}^i x_j = a_{i,\,m+1}^i, \quad i = 1, \ldots, k,$$

$$\sum_{j=k+1}^{m} a_{ij}^k x_j = a_{i,\,m+1}^k, \quad i = k+1, \ldots, m.$$

Cherchons un l tel que $|a_{k+1,\,l}^k| = \max_j |a_{k+1,\,j}^k|$ et notons $x_{k+1} =$ $= x_l$ et $x_l = x_{k+1}$; éliminons l'inconnue x_{k+1} entre toutes les équations à commencer par la $k+2$-ième. Ce changement de notation modifie l'ordre d'élimination et réduit notablement, dans beaucoup de cas, la sensibilité de la solution aux arrondis.

La plupart des méthodes directes de résolution des systèmes d'équations linéaires sont en fait autant de variantes de la méthode

de Gauss qui ne diffèrent les unes des autres que par des détails. Elles impliquent à peu près le même nombre d'opérations arithmétiques, cependant certaines mises en œuvre encombrent moins la mémoire de la machine. Ainsi, avec une mémoire à N cellules, on peut résoudre par les *méthodes d'élimination et d'ascension optimales* des systèmes de l'ordre de $\sim 2\sqrt{N}$. La *méthode de réflexion* qui se distingue un peu de l'algorithme de Gauss, offre l'avantage d'un schéma de calcul homogène.

On demande souvent de résoudre plusieurs systèmes d'équations $A\mathbf{x} = \mathbf{b}_q$, $q = 1, \ldots, p$, ayant une même matrice A. On procède comme suit. Introduisons la notation

$$\mathbf{b}_q = (a_{1,\, m+q}, \ldots, a_{m,\, m+q})^T,$$

et utilisons les formules (4) en calculant d_{ik} pour $i < k \leqslant m + p$.

On aboutit à p systèmes d'équations de matrice triangulaire qui correspondent au problème initial

$$D\mathbf{x} = \mathbf{d}_q, \quad \mathbf{d}_q = (d_{1,\, m+q}, \ldots, d_{m,\, m+q})^T, \quad q = 1, \ldots, p;$$

résolvons ces systèmes un par un. Le nombre total d'opérations arithmétiques nécessaires pour résoudre p systèmes d'équations par ce procédé est $\sim \frac{2}{3}m^3 + 2pm^2$.

On se sert de ce procédé pour se faire entre autres une idée de l'erreur d'arrondi sur la solution sans pour autant consentir d'importantes dépenses supplémentaires. On se donne un vecteur $\mathbf{z}$ de composantes ayant si possible le même ordre et le même signe que celles de la solution cherchée; le manque d'information fait souvent prendre $\mathbf{z} = (1, \ldots, 1)^T$. On calcule le vecteur $\mathbf{c} = A\mathbf{z}$ et on résout simultanément le système initial et le système $A\mathbf{z} = \mathbf{c}$.

Soient $\mathbf{x}'$ et $\mathbf{z}'$ deux solutions réelles de ces systèmes. On pourrait se représenter ce qu'est l'erreur $\mathbf{x}' - \mathbf{x}$ sur la solution cherchée si l'on admet que les erreurs relatives provenant de la résolution par élimination de systèmes ayant une même matrice et des seconds membres distincts, à savoir

$$\frac{\|\mathbf{x}-\mathbf{x}'\|}{\|\mathbf{x}'\|} \quad \text{et} \quad \frac{\|\mathbf{z}-\mathbf{z}'\|}{\|\mathbf{z}'\|},$$

diffèrent d'un nombre de fois pas très grand.

Il existe un autre procédé permettant de se faire une idée de la grandeur réelle de l'erreur d'arrondi. Il s'agit du *changement d'échelle* qui modifie l'allure de l'accumulation de l'erreur de calcul. On résout par cette méthode le système initial et le système

$$(\alpha A)\,\mathbf{x}' = \beta\mathbf{b}, \quad \alpha \text{ et } \beta \text{ nombres};$$

pour α et β des puissances non entières de 2, la comparaison des vecteurs $\mathbf{x}$ et $\frac{\alpha}{\beta}\,\mathbf{x}'$ permet de juger de l'erreur de calcul. On peut par exemple prendre $\alpha = \beta = \sqrt{2}$.

En cas de systèmes d'équations linéaires de matrice hermitienne A il est avantageux de recourir à la *méthode de la racine carrée*. La ma-

trice A est représentée par

$$A = S^*DS, \tag{6}$$

où S est une matrice triangulaire droite et S^* sa conjuguée:

$$S = \begin{pmatrix} s_{11} & s_{12} & \dots \\ 0 & s_{22} & \dots \\ \dots & \dots & \dots \end{pmatrix}$$

avec tous les s_{ii} supérieurs à zéro et D une matrice diagonale telle que d_{ii} valent $+1$ ou -1. L'égalité matricielle (6) forme un système d'équations

$$a_{ij} = \bar{s}_{1i}s_{1j}d_{11} + \dots + \bar{s}_{ii}s_{ij}d_{ii} \quad \text{pour} \quad i \leqslant j.$$

On a négligé les équations analogues pour $i > j$, étant donnée l'équivalence des équations correspondant aux couples (i, j) et (j, i). D'où les relations de récurrence pour déterminer d_{ii} et s_{ij}:

$$d_{ii} = \operatorname{sg}\left(a_{ii} - \sum_{k=1}^{i-1} |s_{ki}|^2 d_{kk}\right),$$

$$s_{ii} = \sqrt{\left|a_{ii} - \sum_{k=1}^{i-1} |s_{ki}|^2 d_{kk}\right|},$$

$$s_{ij} = \frac{a_{ij} - \sum_{k=1}^{i-1} \bar{s}_{ki}s_{kj}d_{kk}}{\bar{s}_{ii}d_{ii}} \quad \text{pour } i < j.$$

La matrice S est triangulaire droite, c'est pourquoi, une fois qu'on a obtenu la représentation (6), la résolution du système initial se ramène à résoudre successivement deux systèmes de matrices triangulaires. Notons que pour $A > 0$ tous les $d_{ii} = 1$ et $A = S^*S$.

Si l'on a des raisons de penser que la solution obtenue $\mathbf{x}^1$ est fortement influencée par l'erreur de calcul, on procède comme suit. Définissons le vecteur $\mathbf{b}^1 = \mathbf{b} - A\mathbf{x}^1$. L'erreur $\mathbf{r}^1 = \mathbf{x} - \mathbf{x}^1$ vérifie évidemment le système d'équations

$$A\mathbf{r}^1 = A\mathbf{x} - A\mathbf{x}^1 = \mathbf{b}^1. \tag{7}$$

En le résolvant avec arrondi nous obtenons $\mathbf{r}^{(1)}$, une valeur approchée de $\mathbf{r}^1$. Posons $\mathbf{x}^2 = \mathbf{x}^1 + \mathbf{r}^{(1)}$. Si la première approximation n'est pas suffisamment précise, répétons l'opération. En résolvant (7), les termes du second membre subissent les mêmes opérations linéaires que ceux à droite dans (1). Dans le calcul sur machine à virgule flottante, il fait donc s'attendre à ce que les solutions de ces systèmes soient entachées d'erreurs relatives d'un même ordre de grandeur. Les erreurs d'arrondi étant en général faibles, on a $\| \mathbf{b}^1 \| \ll \| \mathbf{b} \|$; alors

$\| \mathbf{r}^1 \| \ll \| \mathbf{x}^1 \|$, et la solution de (7) se définit apparemment avec une erreur absolue sensiblement inférieure à celle faite sur la solution de (1). Ainsi, le procédé décrit augmente la précision de la solution approchée.

Le procédé en question est particulièrement commode si les matrices B et D sont logées en mémoire tout au long des calculs. Chaque dégrossissage de la solution implique la recherche du vecteur $\mathbf{b}^k = \mathbf{b} - A\mathbf{x}^k$ et la résolution de deux systèmes de matrices triangulaires, soit $\sim 4m^2$ opérations arithmétiques, ce qui est peu en regard des $\sim 2/3\ m^3$ opérations que coûte la représentation de A sous la forme $A = BD$.

La technique de dégrossissage de la solution qui vient d'être décrite admet souvent une mise en œuvre suivante. Soit B une matrice proche dans un certain sens de la matrice A, mais telle que la résolution du système $B\mathbf{x} = \mathbf{c}$ exige un volume de calcul essentiellement moindre que celle du système $A\mathbf{x} = \mathbf{b}$. Acceptons la solution de $B\mathbf{x} = \mathbf{b}$ pour première approximation $\mathbf{x}^1$ de la solution cherchée. La différence $\mathbf{x} - \mathbf{x}^1$ vérifie le système d'équations

$$A(\mathbf{x} - \mathbf{x}^1) = \mathbf{b} - A\mathbf{x}^1.$$

Au lieu de le résoudre, trouvons la solution de

$$B\mathbf{r}^1 = \mathbf{b} - A\mathbf{x}^1$$

et posons $\mathbf{x}^2 = \mathbf{x}^1 + \mathbf{r}^1$. Chaque valeur approchée s'obtient donc de la précédente par la formule

$$\mathbf{x}^{n+1} = \mathbf{x}^n + B^{-1}(\mathbf{b} - A\mathbf{x}^n) = (E - B^{-1}A)\mathbf{x}^n + B^{-1}\mathbf{b}.$$

Si les matrices A et B sont assez proches, la matrice $E - B^{-1}A$ est munie d'une petite norme, et un tel processus itératif converge rapidement (v. § 3).

L'inversion de matrices est un problème que l'on rencontre bien plus rarement que la résolution de systèmes d'équations. Pour la matrice inverse $X = A^{-1}$, on a $AX = BDX = E$. Il suffit, pour trouver X, de résoudre successivement deux systèmes matriciels $BY = E$, $DX = Y$. On établit sans peine que le calcul de la matrice A^{-1} coûte en gros $2m^3$ opérations arithmétiques.

Si le besoin s'en fait sentir, on améliore les approximations à l'aide du processus itératif $X_k = X_{k-1}(2E - AX_{k-1})$. Pour en établir la convergence considérons la matrice $G_k = E - AX_k$. On a

$$G_k = E - AX_k = E - AX_{k-1}(2E - AX_{k-1}) = \\ = (E - AX_{k-1})^2 = G_{k-1}^2.$$

Il en résulte la suite d'égalités

$$G_k = G_{k-1}^2 = G_{k-2}^4 = \ldots = G_0^{2^k}.$$

Puisque

$$A^{-1} - X_k = A^{-1}(E - AX_k) = A^{-1}G_k = A^{-1}G_0^{2^k},$$

on a l'estimation

$$\| A^{-1} - X_k \| \leqslant \| A^{-1} \| \ \| G_0 \|^{2^k};$$

par conséquent, pour une approximation initiale suffisamment bonne, i.e. si $\| E - AX_0 \| < 1$, le processus itératif converge rapidement.

Dans le cas de systèmes d'ordre élevé, les quantités nécessaires ne peuvent s'implanter simultanément en mémoire rapide et le problème se pose de les loger dans des parties distinctes de la mémoire et d'organiser le travail de façon que le transfert de ces quantités ralentisse le moins possible la marche des calculs. On s'adresse alors à d'autres variantes de la méthode d'élimination telles que les méthodes par bloc. Nous ne nous attarderons pas sur ces problèmes importants pour la bonne raison que la solution optimale du problème considéré dépend souvent essentiellement de la machine utilisée.

§ 2. Méthode d'orthogonalisation

Ecrivons le système $Ax = b$ sous la forme

$$a_{11}x_1 + \ldots + a_{1m}x_m + a_{1,\,m+1} = 0,$$
$$\ldots\ldots\ldots\ldots\ldots\ldots$$
$$a_{m1}x_1 + \ldots + a_{mm}x_m + a_{m,\,m+1} = 0.$$

Une autre écriture de ces équations est $(\mathbf{a}_1, \mathbf{y}) = 0, \ldots, (\mathbf{a}_m, \mathbf{y}) = 0$, où $\mathbf{a}_i = (a_{i1}, \ldots, a_{i,\,m+1})$, $\mathbf{y} = (x_1, \ldots, x_m, 1)$. Résoudre le système initial équivaut donc à trouver le vecteur $\mathbf{y}$ de dernière composante égale à 1 et orthogonal aux vecteurs $\mathbf{a}_1, \ldots, \mathbf{a}_m$. Adjoignons à ce système le vecteur $\mathbf{a}_{m+1} = (0, \ldots, 0, 1)$. Les vecteurs $\mathbf{a}_1, \ldots, \mathbf{a}_{m+1}$ sont linéairement indépendants puisque le déterminant de la matrice dont ils constituent les lignes est égal à celui de la matrice A qu'on suppose non dégénérée. Appliquons au système $(\mathbf{a}_1, \ldots, \mathbf{a}_{m+1})$ l'algorithme d'orthogonalisation successive pour la norme $\| \mathbf{u} \| = \sqrt{(\mathbf{u}, \mathbf{u})}$. Pour construire les vecteurs $\mathbf{u}_1, \mathbf{v}_1, \mathbf{u}_2, \mathbf{v}_2, \ldots, \mathbf{u}_{m+1}, \mathbf{v}_{m+1}$ on utilise la règle habituelle $\mathbf{u}_1 = \mathbf{a}_1$, $\mathbf{v}_1 = \dfrac{\mathbf{u}_1}{\| \mathbf{u}_1 \|}$, $\ldots$, $\mathbf{u}_k = \mathbf{a}_k - \sum\limits_{i=1}^{k-1} c_{ki}\mathbf{v}_i$, $\mathbf{v}_k = \dfrac{\mathbf{u}_k}{\| \mathbf{u}_k \|}$, $\ldots$ L'orthogonalité de $\mathbf{u}_k$ aux $\mathbf{v}_1, \ldots, \mathbf{v}_{k-1}$ donne les équations $0 = (\mathbf{u}_k, \mathbf{v}_i) = (\mathbf{a}_k, \mathbf{v}_i) - c_{ki}$, i.e. $c_{ki} = (\mathbf{a}_k, \mathbf{v}_i)$ pour $i < k$. Donc,

$$\mathbf{u}_k = \mathbf{a}_k - \sum_{i=1}^{k-1} (\mathbf{a}_k, \mathbf{v}_i)\,\mathbf{v}_i.$$

La dernière coordonnée z_{m+1} du vecteur $\mathbf{u}_{m+1} = (z_1, \ldots, z_{m+1})$ est différente de zéro ; en supposant le contraire, nous aurions obtenu que les conditions

$$(\mathbf{u}_{m+1}, \mathbf{a}_i) = 0, \quad i = 1, \ldots, m,$$

équivalentes aux conditions $(\mathbf{u}_{m+1}, \mathbf{v}_i) = 0$, $i = 1, \ldots, m$, forment un système d'équations linéaires $\sum\limits_{j=1}^{m} a_{ij}z_j = 0$, $i = 1, \ldots, m$, admettant une solution non nulle ; alors le déterminant du système de départ devrait être nul.

Ainsi, les conditions $(\mathbf{u}_{m+1}, \mathbf{a}_i) = 0$, $i = 1, \ldots, m$, peuvent s'écrire

$$a_{i1}\frac{z_1}{z_{m+1}} + \ldots + a_{im}\frac{z_m}{z_{m+1}} + a_{i,\,m+1} = 0 ;$$

il s'ensuit que $\left(\dfrac{z_1}{z_{m+1}}, \ldots, \dfrac{z_m}{z_{m+1}}\right)$ est solution du système initial.

Cette variante de la méthode d'orthogonalisation coûte $\sim m^3$ multiplications et divisions et $\sim m^3$ additions et cède en précision à la méthode de Gauss. Essayons d'améliorer le résultat. Cherchons le vecteur $\mathbf{y} = (y_1, \ldots, y_m, 1)$ satisfaisant aux conditions $(\mathbf{a}_i, \mathbf{y}) = 0$, $i = 1, \ldots, m$, sous la forme

$$\mathbf{y} = \sum_{j=1}^{m+1} d_j \mathbf{v}_j.$$

Les relations $(\mathbf{a}_i, \mathbf{y}) = 0$ donnent le système d'équations

$$\sum_{j=1}^{m+1} d_j (\mathbf{a}_i, \mathbf{v}_j) = 0, \quad i = 1, \ldots, m;$$

de plus, la condition d'égalité de 1 et de la $(m+1)$-ième coordonnée du vecteur $\mathbf{y}$ fournit

$$\sum_{j=1}^{m+1} d_j (\mathbf{v}_j)_{m+1} = 1;$$

ici et dans ce qui suit $(v)_k$ est la k-ième coordonnée de $\mathbf{v}$. En calculant sans arrondi, les propriétés de la famille de vecteurs construite font que $(\mathbf{a}_i, \mathbf{v}_j) = 0$, $i < j$, et que le système considéré s'écrit

$$\left.\begin{aligned} & d_1 (\mathbf{a}_1, \mathbf{v}_1) && && = 0, \\ & d_1 (\mathbf{a}_2, \mathbf{v}_1) + d_2 (\mathbf{a}_2, \mathbf{v}_2) && && = 0, \\ & d_1 (\mathbf{a}_m, \mathbf{v}_1) + \ldots && + d_m (\mathbf{a}_m, \mathbf{v}_m) && = 0, \\ & d_1 (\mathbf{v}_1)_{m+1} + \ldots && && + d_{m+1} (\mathbf{v}_{m+1})_{m+1} = 1, \end{aligned}\right\} \quad (1)$$

La solution de ce système est le vecteur $\mathbf{d}^0 = (0, \ldots, 0, ((\mathbf{v}_{m+1})_{m+1})^{-1})$.

En arrondissant, la matrice du système (1) aura des éléments petits non nuls au-dessus de la diagonale. Ecrivons le système obtenu sous la forme $C\mathbf{d} + D\mathbf{d} = \mathbf{g}$, $\mathbf{g} = (0, \ldots, 0, 1)^T$, C étant la matrice du système (1) et

$$D = \begin{pmatrix} 0 & (\mathbf{a}_1, \mathbf{v}_2) & (\mathbf{a}_1, \mathbf{v}_3) & \ldots & (\mathbf{a}_1, \mathbf{v}_{m+1}) \\ 0 & 0 & (\mathbf{a}_2, \mathbf{v}_3) & \ldots & (\mathbf{a}_2, \mathbf{v}_{m+1}) \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & 0 & \ldots & 0 \end{pmatrix}.$$

Pour l'approximation initiale $\mathbf{d}^0$ donnée, déterminons les valeurs améliorées $\mathbf{d}^k$ à l'aide du processus itératif

$$C\mathbf{d}^{k+1} + D\mathbf{d}^k = \mathbf{g}. \quad (2)$$

La relation (2) s'écrit

$$\mathbf{d}^{k+1} = -C^{-1}D\mathbf{d}^k + C^{-1}\mathbf{g}.$$

La petitesse des éléments de D entraîne celle de la norme de la matrice $-C^{-1}D$; aussi (v. § 3) la suite d'approximations obtenues par itérations converge très rapidement. Ce procédé atténue la répercussion de l'erreur de calcul. Pour se prémunir davantage contre l'influence notable de celle-ci, le procédé est répété à chaque pas d'orthogonalisation. En présence d'arrondis il se peut que $(\mathbf{v}_i, \mathbf{v}_j) \neq 0$, $i \neq j$; c'est pourqu i les conditions $(\mathbf{u}_k, \mathbf{v}_j) = 0$, $j < k$, forment le système

$$(\mathbf{a}_k, \mathbf{v}_j) - \sum_{i=1}^{k-1} c_{ki} (\mathbf{v}_i, \mathbf{v}_j) = 0$$

de matrice proche d'une matrice diagonale. Les valeurs c_{ki} s'obtiennent par itérations du type (2).

§ 3. Méthode itérative simple

Le système d'équations

$$A\mathbf{x} = \mathbf{b} \tag{1}$$

se transforme en

$$\mathbf{x} = B\mathbf{x} + \mathbf{c}, \tag{2}$$

et sa solution est cherchée comme la limite de la suite

$$\mathbf{x}^{n+1} = B\mathbf{x}^n + \mathbf{c}. \tag{3}$$

Il est évident que tout système

$$\mathbf{x} = \mathbf{x} - D\,(A\mathbf{x} - \mathbf{b}) \tag{4}$$

s'écrit comme (2) et est équivalent à (1) pour $\det D \neq 0$. D'autre part, chaque système (2) équivalent à (1) s'écrit sous forme (4) de matrice $D = (E - B)A^{-1}$.

Récapitulons en quelques mots les définitions des normes principales dans les espaces de vecteurs et de matrices. Si l'espace vectoriel $\mathbf{x}\,(x_1, \ldots, x_m)^T$ est muni de la norme $\|\mathbf{x}\|$, la norme qui lui est *associée* dans l'espace des matrices A est

$$\|A\| = \sup_{x \neq 0} \frac{\|Ax\|}{\|\mathbf{x}\|}.$$

Voici les normes les plus usuelles :

$$\|\mathbf{x}\|_1 = \max_{1 \leqslant j \leqslant m} |x_j|,$$

$$\|\mathbf{x}\|_2 = \sum_{j=1}^{m} |x_j|,$$

$$\|\mathbf{x}\|_3 = \sqrt{\sum_{j=1}^{m} |x_j|^2} = \sqrt{(\mathbf{x}, \mathbf{x})}.$$

Les normes qui leur sont associées dans l'espace de matrices sont respectivement

$$\|A\|_1 = \max_{1 \leqslant i \leqslant m} \Big(\sum_{j=1}^{m} |a_{ij}|\Big), \tag{5}$$

$$\|A\|_2 = \max_{1 \leqslant j \leqslant m} \Big(\sum_{i=1}^{m} |a_{ij}|\Big), \tag{6}$$

$$\|A\|_3 = \sqrt{\max_{1 \leqslant i \leqslant m} \lambda^i_{A^*A}}\,; \tag{7}$$

ici et dans la suite λ^i_D désignent les valeurs propres de la matrice D.

Déduisons ces relations. On a

$$\|Ax\|_1 = \max_i \left|\sum_j a_{ij}x_j\right| \leqslant$$

$$\leqslant \max_i \left(\sum_j |a_{ij}| \max_j |x_j|\right) \leqslant \max_i \left(\sum_j |a_{ij}|\right) \max_j |x_j|,$$

c'est-à-dire

$$\frac{\|Ax\|_1}{\|x\|_1} \leqslant \max_i \left(\sum_j |a_{ij}|\right).$$

Supposons que $\max_i \left(\sum_j |a_{ij}|\right)$ est atteint pour $i = l$; pour le vecteur

$$x = (\operatorname{sg}(a_{l1}), \ldots, \operatorname{sg}(a_{lm}))^T$$

on a

$$\|x\|_1 = 1,$$

$$\|Ax\|_1 \geqslant \left|\sum_j a_{lj}x_j\right| = \sum_j |a_{lj}| = \left(\max_i \sum_j |a_{ij}|\right) \|x\|_1.$$

D'où (5).

De même

$$\|Ax\|_2 = \sum_i \left|\sum_j a_{ij}x_j\right| \leqslant \sum_i \sum_j |a_{ij}||x_j| =$$

$$= \sum_j \left(\sum_i |a_{ij}|\right) |x_j| \leqslant \left(\max_j \sum_i |a_{ij}|\right) \sum_j |x_j|,$$

i.e.

$$\frac{\|Ax\|_2}{\|x\|_2} \leqslant \max_j \left(\sum_i |a_{ij}|\right).$$

Supposons que $\max_j \sum_i |a_{ij}|$ est réalisé pour $j = l$. On a, dans le cas du vecteur x dont toutes les composantes sont nulles sauf x_l,

$$\|Ax\|_2 = \sum_i \left|\sum_j a_{ij}x_j\right| = \sum_i |a_{il}||x_l| =$$

$$= \left(\sum_i |a_{il}|\right) |x_l| = \left(\max_j \sum_i |a_{ij}|\right) \sum_j |x_j| = \left(\max_j \sum_i |a_{ij}|\right) \|x\|_2,$$

ce qui entraîne (6).

Par définition

$$\|A\|_3 = \sup_x \frac{\|Ax\|_3}{\|x\|_3} = \sup_x \sqrt{\frac{(Ax, Ax)}{(x, x)}} = \sqrt{\sup_x \frac{(A^*Ax, x)}{(x, x)}}.$$

La matrice A^*A est symétrique vu que $(A^*A)^* = A^*(A^*)^* = A^*A$.

Soient B une matrice symétrique, $e_1, \ldots, e_m$ un système orthonormé de ses vecteurs propres et $\lambda_1, \ldots, \lambda_m$ les valeurs propres

correspondantes. Tout vecteur $\mathbf{x}$ peut s'écrire sous la forme $\sum_i c_i \mathbf{e}_i$. Evidemment

$$(B\mathbf{x}, \mathbf{x}) = \left(\sum_i \lambda_i c_i \mathbf{e}_i, \sum_i c_i \mathbf{e}_i\right) = \sum_i \lambda_i |c_i|^2;$$

aussi

$$(B\mathbf{x}, \mathbf{x}) \leqslant (\max_i \lambda_i) \sum_i |c_i|^2 = (\max_i \lambda_i)(\mathbf{x}, \mathbf{x}) \tag{8}$$

et

$$(B\mathbf{x}, \mathbf{x}) \geqslant (\min_i \lambda_i)(\mathbf{x}, \mathbf{x}). \tag{9}$$

D'autre part, $\frac{(B\mathbf{e}_i, \mathbf{e}_i)}{(\mathbf{e}_i, \mathbf{e}_i)} = \lambda_i$. Ces relations entraînent

$$\sup_{\mathbf{x}} \frac{|(B\mathbf{x}, \mathbf{x})|}{(\mathbf{x}, \mathbf{x})} = \max_i |\lambda_i|. \tag{10}$$

Puisque $(A^*A\mathbf{x}, \mathbf{x}) = (A\mathbf{x}, A\mathbf{x}) \geqslant 0$, tous les $\lambda^i_{A^*A} \geqslant 0$; posant $B = A^*A$ dans (10), il vient

$$\sup_{\mathbf{x}} \frac{|(A^*A\mathbf{x}, \mathbf{x})|}{(\mathbf{x}, \mathbf{x})} = \max_i |\lambda^i_{A^*A}| = \max_i \lambda^i_{A^*A}.$$

Les relations obtenues impliquent (7).

Indiquons un cas particulier important. Si A est symétrique, alors $\lambda^i_{A^*A} = \lambda^i_{A^2} = |\lambda^i_A|^2$; donc pour une matrice symétrique

$$\|A\|_3 = \max_i |\lambda^i_A|. \tag{11}$$

Si $A\mathbf{x} = \lambda\mathbf{x}$, alors $\|A\| \, \|\mathbf{x}\| \geqslant \|\lambda\mathbf{x}\| = |\lambda| \, \|\mathbf{x}\|$; ainsi le module de toute valeur propre de la matrice A est au plus égal à sa norme.

THÉORÈME (sur la condition suffisante de convergence de la méthode itérative simple). *Si* $\|B\| < 1$, *le système d'équations* (2) *possède une solution et une seule et le processus itératif* (3) *converge vers la solution en progression géométrique.*

DÉMONSTRATION. On a, pour toute solution du système (2), $\|\mathbf{x}\| \leqslant \|B\| \, \|\mathbf{x}\| + \|\mathbf{c}\|$; d'où $\|\mathbf{x}\|(1 - \|B\|) \leqslant \|\mathbf{c}\|$ ou $\|\mathbf{x}\| \leqslant (1 - \|B\|)^{-1} \|\mathbf{c}\|$. Cette inégalité entraîne l'existence et l'unicité de la solution du système homogène $\mathbf{x} = B\mathbf{x}$, donc du système (2). Soit X la solution de (2). (2) et (3) donnent une équation par rapport à l'erreur $\mathbf{r}^n = \mathbf{x}^n - \mathbf{X}$:

$$\mathbf{r}^{n+1} = B\mathbf{r}^n. \tag{12}$$

On a immédiatement

$$\mathbf{r}^n = B^n\mathbf{r}^0. \tag{13}$$

Il découle de (13) que $\| \mathbf{r}^n \| \leqslant \| B \|^n \| \mathbf{r}_0 \| \to 0$. Le théorème est démontré.

La qualité du processus itératif est bien caractérisée par la vitesse de décroissance du rapport de l'erreur commise au bout de n itérations à l'erreur d'approximation initiale

$$s_n = \sup_{\mathbf{x}^0 \neq \mathbf{x}} \frac{\| \mathbf{r}^n \|}{\| \mathbf{r}^0 \|} = \sup_{\mathbf{r}^0 \neq 0} \frac{\| B^n \mathbf{r}^0 \|}{\| \mathbf{r}^0 \|} = \| B^n \|.$$

On garantit que $s_n \leqslant \varepsilon$ si $\| B \|^n \leqslant \varepsilon$, i.e.

$$n \geqslant n_\varepsilon = \ln (\varepsilon^{-1}) / \ln (\| B \|^{-1}). \tag{14}$$

S'il existe des $\gamma_{\alpha\beta}$, $\gamma_{\beta\alpha}$ constants tels que pour $\mathbf{x} \neq \mathbf{0}$

$$\frac{\| \mathbf{x} \|_\beta}{\| \mathbf{x} \|_\alpha} \leqslant \gamma_{\alpha\beta}, \quad \frac{\| \mathbf{x} \|_\alpha}{\| \mathbf{x} \|_\beta} \leqslant \gamma_{\beta\alpha},$$

les normes $\| \mathbf{x} \|_\alpha$ et $\| \mathbf{x} \|_\beta$ sont dites *équivalentes*. On a l'inégalité évidente

$$\| \mathbf{r}^n \|_\beta \leqslant \gamma_{\alpha\beta} \| \mathbf{r}^n \|_\alpha \leqslant \gamma_{\alpha\beta} \| B \|_\alpha^n \| \mathbf{r}^0 \|_\alpha \leqslant \gamma_{\alpha\beta} \gamma_{\beta\alpha} \| B \|_\alpha^n \| \mathbf{r}^0 \|_\beta.$$

Ainsi, si la condition du théorème démontré est réalisée pour une certaine norme, le théorème est valable pour toute norme équivalente.

Deux normes quelconques dans un espace de dimension finie sont équivalentes ; il en est en particulier de $\| \mathbf{x} \|_1$, $\| \mathbf{x} \|_2$, $\| \mathbf{x} \|_3$ vu la chaîne d'inégalités évidentes

$$\| \mathbf{x} \|_1 \leqslant \| \mathbf{x} \|_3 \leqslant \| \mathbf{x} \|_2 \leqslant m \| \mathbf{x} \|_1.$$

LEMME. *Supposons que toutes les valeurs propres λ_i d'une matrice B appartiennent au cercle $|\lambda| \leqslant q$ et qu'aux λ_i de module q correspondent des blocs de Jordan de dimension 1. Il existe alors une matrice $\Lambda = D^{-1}BD$ munie de la norme $\| \Lambda \|_1 \leqslant q$.*

DÉMONSTRATION. Posons $\eta = q - \max_{|\lambda_i| < q} |\lambda_i|$. Les valeurs propres de la matrice $\eta^{-1}B$ sont $\eta^{-1}\lambda_i$. Transformons la matrice $\eta^{-1}B$ en forme de Jordan

$$D^{-1}(\eta^{-1}B)D = \begin{pmatrix} \eta^{-1}\lambda_1 & \alpha_{12} & 0 & \dots \\ 0 & \eta^{-1}\lambda_2 & \alpha_{23} & \dots \\ \dots & \dots & \dots & \dots \end{pmatrix}$$

avec $\alpha_{i,\,i+1}$ pouvant être 0 ou 1.

La multiplication par η donne

$$\Lambda = D^{-1}BD = \begin{pmatrix} \lambda_1 & \alpha_{12}\eta & 0 & \dots \\ 0 & \lambda_2 & \alpha_{23}\eta & \dots \\ \dots & \dots & \dots & \dots \end{pmatrix}.$$

Si $|\lambda_i| = q$, alors, selon la condition du lemme, $\alpha_{i,i+1} = 0$, donc $|\lambda_i| + |\alpha_{i,i+1}\eta| = q$; si $|\lambda_i| < q$, on a

$$|\lambda_i| + |\alpha_{i,i+1}\eta| \leqslant \max_{|\lambda_i|<q} |\lambda_i| + \eta = q.$$

Ainsi, $\| \Lambda \|_1 = \max_i (|\lambda_i| + |\alpha_{i,i+1}\eta|) \leqslant q$.

Nous aurons besoin de ce lemme dans le chapitre VIII. Son cas particulier nous permettra de démontrer le

THÉORÈME (sur la condition nécessaire et suffisante de convergence de la méthode itérative simple). *Supposons que le système* (2) *admet une solution unique. Le processus itératif* (3) *converge vers la solution de* (2) *quelle que soit l'approximation initiale si et seulement si toutes les valeurs propres de la matrice B ont un module plus petit que* 1.

SUFFISANCE. Soit un q quelconque dans les limites $\max_i |\lambda_i| < q < 1$. La condition du lemme ci-dessus est vérifiée pour ce q, et il existe donc une matrice D telle que $\| \Lambda \|_1 \leqslant q$ pour $\Lambda = D^{-1}BD$. Puisque $B = D\Lambda D^{-1}$, il vient

$$B^n = D\Lambda D^{-1}D \ldots D^{-1}D\Lambda D^{-1} = D\Lambda^n D^{-1}.$$

Aussi

$$\| B^n \|_1 \leqslant \| D \|_1 \| D^{-1} \|_1 q^n \to 0$$

et

$$\| \mathbf{x}^n - \mathbf{X} \|_1 \leqslant \| D \|_1 \| D^{-1} \|_1 q^n \| \mathbf{x}^0 - \mathbf{X} \|_1 \to 0 \qquad (15)$$

avec n tendant vers ∞.

Si χ_i sont les vecteurs unités dirigés suivant les axes de coordonnées, $\mathbf{x} = (x_1, \ldots, x_m)^T$, alors $\mathbf{x} = \sum_i x_i \chi_i$ et

$$\| \mathbf{x} \| \leqslant \sum_i |x_i| \| \chi_i \| \leqslant \left(\sum_i \| \chi_i \|\right) \| \mathbf{x} \|_1.$$

On a donc, quelle que soit la norme,

$$\| \mathbf{x}^n - \mathbf{X} \| \leqslant \left(\sum_i \| \chi_i \|\right) \| D \|_1 \| D^{-1} \|_1 q^n \| \mathbf{x}^0 - \mathbf{X} \|_1 \to 0. \qquad (16)$$

Les relations (15) et (16) signifient également que toute norme de l'erreur décroît plus vite que n'importe quelle progression géométrique de raison supérieure à $\max_i |\lambda_i|$.

NÉCESSITÉ. Soient $|\lambda_l| \geqslant 1$ et $\mathbf{e}_l$ le vecteur propre correspondant de la matrice B. Alors on a, pour l'approximation de départ $\mathbf{x}^0 = \mathbf{X} + c\mathbf{e}_l$, $c \neq 0$:

$$\mathbf{r}^0 = c\mathbf{e}_l \quad \text{et} \quad \mathbf{r}^n = \lambda_l^n c\mathbf{e}_l \not\to \mathbf{0}, \quad n \to \infty.$$

PROBLÈME. Supposons que toutes les valeurs propres de la matrice B, sauf $\lambda_1 = 1$ simple, sont intérieures au cercle unité et que le système (2) a X pour solution; tous les $x = X + ce_1$ sont visiblement solution de (2). Démontrer que le processus itératif (3) converge vers l'une de ces solutions.

§ 4. Processus itératif réel

Si toutes les valeurs propres de B appartiennent à l'intérieur du cercle unité, il semblerait qu'on n'a pas à se soucier de la qualité de la méthode travaillant avec des nombres ayant un exposant borné et avec arrondi. En effet, les perturbations d'approximation par suite d'arrondis sont équivalentes aux perturbations des conditions initiales du processus itératif. Du moment que ce dernier est convergent et « autocorrecteur », ces perturbations finissent par s'amortir, et on obtient une bonne approximation de la solution du problème initial.

Or, en résolvant certains systèmes, on a été confronté à une situation différente. Toutes les valeurs propres de la matrice B étaient bel et bien intérieures au cercle $\| \lambda \| \leqslant 1/2$, mais le processus itératif s'arrêtait au bout d'un certain nombre d'itérations à cause du dépassement de capacité. Dans d'autres cas, le dépassement n'avait pas lieu, mais les vecteurs x^n obtenus n'avaient pas tendance à se stabiliser. Ce dernier cas est particulièrement dangereux. Un utilisateur débutant en conclurait que, dans la condition $\max | \lambda_i | \leqslant \leqslant 1/2$, un tel nombre d'itérations, disons 100, suffirait amplement pour obtenir une solution avec la précision demandée. Il effectuerait donc ces cent itérations et accepterait comme parfait le résultat obtenu. On a donc jugé nécessaire de pousser l'étude des itérations et d'introduire de nouvelles notions dans la théorie des opérateurs.

Construisons un exemple qui nous permette de suivre le phénomène en question sous forme explicite. Prenons pour modèle un processus itératif correspondant à la matrice bidiagonale

$$B_0 = \begin{pmatrix} \alpha & \beta & 0 & \dots & 0 \\ 0 & \alpha & \beta & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & \beta \\ 0 & 0 & 0 & \dots & \alpha \end{pmatrix}.$$

B_0 élevée à la puissance n donne la matrice triangulaire:

$$B_0^n = [b_{ij}^{(n)}] = \begin{pmatrix} \alpha^n & C_n^1 \alpha^{n-1}\beta & C_n^2 \alpha^{n-2}\beta^2 & \dots \\ 0 & \alpha^n & C_n^1 \alpha^{n-1}\beta & \dots \\ \dots & \dots & \dots & \dots \end{pmatrix}$$

d'éléments $b_{ij}^{(n)} = C_n^{j-i}\alpha^{n-(j-i)}\beta^{j-i}$. Si $\mathbf{r}^0 = (0, \ldots, 0, 1)^T$, alors

$$\mathbf{r}^n = B_0^n\mathbf{r}^0 = (b_{1m}^{(n)}, \ldots, b_{mm}^{(n)})^T, \quad \|\mathbf{r}^n\|_2 = \sum_{i=1}^{m} |b_{im}^{(n)}|.$$

Pour $n < m$, celle-ci se simplifie :

$$\|\mathbf{r}^n\|_2 = \sum_{i=1}^{m} C_n^{m-i} |\alpha|^{n-(m-i)} |\beta|^{m-i} =$$

$$= \sum_{k=0}^{m-1} C_n^k |\alpha|^{n-k} |\beta|^k = \sum_{k=0}^{n} C_n^k |\alpha|^{n-k} |\beta|^k = (|\alpha| + |\beta|)^n.$$

Considérons le cas $|\alpha| < 1$, $|\alpha| + |\beta| > 1$, $|\beta/(1-\alpha)| < 1$. Soit $\mathbf{c} = \mathbf{c}^0 = (0, \ldots, 0, 1)^T$; on vérifie directement que, avec ce $\mathbf{c}$, la solution du système considéré est

$$\mathbf{X}^0 = \left(\frac{1}{1-\alpha}\left(\frac{\beta}{1-\alpha}\right)^{m-1}, \ldots, \frac{1}{1-\alpha}\right)^T ;$$

évidemment,

$$\|\mathbf{X}^0\|_2 \leqslant \frac{1}{|1-\alpha|} \sum_{k=0}^{\infty} \left|\frac{\beta}{1-\alpha}\right|^k = \frac{1}{|1-\alpha|\left(1-\left|\frac{\beta}{1-\alpha}\right|\right)} = \omega.$$

Pour l'approximation initiale $\mathbf{x}^0 = \mathbf{X}^0 + \mathbf{c}^0$, on a $\mathbf{r}^0 = \mathbf{c}^0$ et, conformément aux constructions ci-dessus,

$$\|\mathbf{r}^n\|_2 = (|\alpha| + |\beta|)^n \quad \text{pour} \quad n < m.$$

Prenons un m tel que $\sigma = \frac{(|\alpha| + |\beta|)^{m-1} - \omega}{m}$ dépasse la capacité de la machine. Les relations obtenues plus haut entraînent $\|\mathbf{x}^{m-1}\|_1 \geqslant \geqslant \frac{\|\mathbf{x}^{m-1}\|_2}{m} \geqslant \frac{\|\mathbf{r}^{m-1}\|_2 - \|\mathbf{X}^0\|_2}{m} \geqslant \sigma$; l'exemple que nous venons de construire jouit donc des propriétés suivantes : la norme de l'approximation de départ est petite, le processus itératif converge si l'on se passe d'arrondi et si l'exposant des nombres est borné ; dans les conditions réelles, le processus s'arrête pour $n = m - 1$ au plus par suite des valeurs trop grandes des composantes d'approximations.

Voici un exemple numérique concret satisfaisant aux conditions imposées dans le cas de la plupart des machines modernes : $\alpha = -1/2$, $\beta = 5/4$, $m = 120$.

Soit $|\beta/(1-\alpha)| > 1$, $\mathbf{c}_1 - \mathbf{c}_2 = \delta\mathbf{c}^0$. Les solutions $\mathbf{x}_1$, $\mathbf{x}_2$ des systèmes $\mathbf{x}_i = B_0\mathbf{x}_i + \mathbf{c}_i$, $i = 1, 2$, vérifient alors les relations

$$\|\mathbf{x}_1 - \mathbf{x}_2\|_1 = \delta \|\mathbf{X}^0\|_1 = |\beta|^{-1} |\beta/(1-\alpha)|^m \delta.$$

Si $|\beta/(1-\alpha)|^m$ est très grand, l'erreur d'arrondi sur le second membre, qui est proportionnelle à c^0, entraînera une erreur inacceptable sur la solution.

Pour $|\beta/(1-\alpha)| > 1$ et m grand, le système s'avère un modèle important de systèmes mal conditionnés (v. § 13).

Examinons un cas réel avec arrondi à chaque pas. Il est fort peu probable que la présence d'arrondis nous évite un dépassement de capacité s'il y en a déjà en leur absence. C'est pourquoi nous nous arrêterons à un cas sans dépassement de capacité. A la place de $\mathbf{x}^n$, on obtient alors $\mathbf{x}^{*n}$ reliés par

$$\mathbf{x}^{*n+1} = B\mathbf{x}^{*n} + \mathbf{c} + \boldsymbol{\rho}^n,$$

$\boldsymbol{\rho}^n$ étant l'arrondi total par pas d'itération.

Il en résulte, compte tenu de (3.2), l'équation relativement à l'erreur réelle $\mathbf{r}^{*n} = \mathbf{x}^{*n} - \mathbf{X}$:

$$\mathbf{r}^{*n+1} = \boldsymbol{\rho}^n + B\mathbf{r}^{*n}. \tag{1}$$

Exprimant chaque $\mathbf{r}^{*n}$ par l'itérée précédente, il vient

$$\mathbf{r}^{*n} = \boldsymbol{\rho}^{n-1} + B\mathbf{r}^{*n-1} = \boldsymbol{\rho}^{n-1} + B(\boldsymbol{\rho}^{n-2} + B\mathbf{r}^{*n-2}) =$$
$$= \boldsymbol{\rho}^{n-1} + B\boldsymbol{\rho}^{n-2} + \ldots + B^{n-1}\boldsymbol{\rho}^0 + B^n\mathbf{r}^0. \tag{2}$$

Pour $|\alpha| < 1$, $|\alpha| + |\beta| > 1$, la norme $\| B_0^n \|$ a, on l'a vu, l'allure suivante: pour n petits elle a tendance à s'accroître, et pour n grands, elle tend vers zéro. On montre que la valeur maximale $\varphi(B_0) = \max_n \| B_0^n \|$ est atteinte pour un $n = n_0$ de l'ordre de m. Avec les normes $\| B^n \|$ se comportant de cette façon on se place dans la situation suivante. La quantité $\max_n \| \mathbf{x}^{*n} \|$ n'est pas si importante pour entraîner un dépassement de capacité et l'arrêt de la machine; d'autre part, $\varphi(B)\, 2^{-t} \gg \omega$, ω étant la plus grande erreur admise sur la solution. Aussi, avec $n > n_0$, il existe toujours, dans le second membre de (2), un terme $B^{n_0}\boldsymbol{\rho}^{n-1-n_0}$ muni d'une norme $\gg\omega$. Résultat: les approximations $\mathbf{x}^{*n}$ ne se stabilisent pas avec une précision acceptable.

L'erreur de calcul totale $\boldsymbol{\rho}_n = \sum_{j=0}^{n-1} B^{n-1-j}\boldsymbol{\rho}^j$ peut s'avérer notable non seulement à cause de la valeur importante de termes isolés, mais aussi parce qu'ils sont nombreux.

Soient B une matrice symétrique, $\| B \|_3 = \max_i |\lambda_B^i| = \lambda_B^1 < 1$ et $\mathbf{e}^1$ le vecteur propre normé associé à λ_B^1. Considérons un modèle d'arrondis tel qu'on ait $\boldsymbol{\rho}^j = \rho\mathbf{e}^1$ à chaque pas, ρ étant de l'ordre de 2^{-t}. On a

$$\boldsymbol{\rho}_n = \rho \sum_{j=0}^{n-1} (\lambda_B^1)^j \mathbf{e}^1 = \rho \frac{1-(\lambda_B^1)^n}{1-\lambda_B^1} \mathbf{e}^1.$$

Puisque le nombre d'itérations est tel que $\| B^n \| \ll 1$ et $\| B^n \| = (\lambda_B^1)^n$, on peut estimer que $\| \rho_n \| \approx \frac{\rho}{1-\lambda_B^1}$. Ainsi, si λ_B^1 est proche de l'unité, l'influence globale des arrondis sur les pas d'itération s'avère assez grande.

Montrons que l'erreur de calcul présente inévitablement cet ordre de grandeur. Supposons qu'au lieu du système (3.2) on demande de résoudre le système $\mathbf{x} = B\mathbf{x} + \mathbf{c} + \rho\mathbf{e}_1$. La différence $\bar{\mathbf{X}} - \mathbf{X}$ des solutions de ces systèmes vérifie $(\bar{\mathbf{X}} - \mathbf{X}) = B(\bar{\mathbf{X}} - \mathbf{X}) + \rho\mathbf{e}_1$; d'où $\bar{\mathbf{X}} - \mathbf{X} = (E - B)^{-1}\rho\mathbf{e}_1 = (1 - \lambda_B^1)^{-1}\rho\mathbf{e}_1$. L'erreur de l'ordre de $(1 - \lambda_B^1)^{-1}\rho$ est donc inévitable et la perturbation des approximations par suite d'itérations est comparable à une erreur inhérente. Afin de réduire le nombre d'itérations lorsqu'on passe au système $\mathbf{x} = B\mathbf{x} + \mathbf{c}$, on tâche d'obtenir un système avec la plus petite $\| B \|$ possible; c'est pourquoi les problèmes avec λ_B supérieur, disons, à 0,9999 sont plutôt rares.

§ 5. Spectre d'une famille de matrices

Soit $\{A_m\}$ une famille de matrices carrées avec m, dimension de la matrice A_m, parcourant un certain sous-ensemble de l'ensemble des nombres naturels, et soient $\| \mathbf{x} \|_{(m)}$ les normes définies dans les espaces de vecteurs $\mathbf{x}_m$ de dimension m.

Appelons $\lambda \neq \infty$ *point du spectre* $S = S(\{A_m\})$ *de la famille de matrices* $\{A_m\}$ sous la condition suivante: pour un $\varepsilon > 0$ quelconque on a une matrice A_m et un vecteur $\mathbf{x}_m \neq \mathbf{0}$ tels que

$$\| A_m\mathbf{x}_m - \lambda\mathbf{x}_m \|_{(m)} \leqslant \varepsilon \| \mathbf{x}_m \|_{(m)}. \tag{1}$$

De même, $\lambda = \infty$ est dit point du spectre de la famille de matrices si, pour tout K, on a une matrice A_m et un vecteur $\mathbf{x}_m \neq \mathbf{0}$ tels que

$$\| A_m\mathbf{x}_m \|_{(m)} \geqslant K \| \mathbf{x}_m \|_{(m)}. \tag{2}$$

L'ensemble des points du spectre constitue le *spectre de la famille de matrices.*

Si la famille $\{A_m\}$ ne comprend qu'une seule matrice A_m, le spectre de la famille coïncide évidemment avec l'ensemble S_m des valeurs propres de A_m.

PROBLÈME. Démontrer que le spectre de toute famille de matrices est un ensemble fermé.

NOTE. De nombreux auteurs définissent autrement le spectre d'une famille de matrices. Le point $\lambda \neq \infty$ est dit *point du spectre d'une famille de matrices* si, pour tout $\varepsilon > 0$, il existe une suite infinie de matrices A_m et de vecteurs $\mathbf{x}_m \neq \mathbf{0}$ associés tels que

$$\| A_m\mathbf{x}_m - \lambda\mathbf{x}_m \|_{(m)} \leqslant \varepsilon \| \mathbf{x}_m \|_{(m)}.$$

Dans le cas $\lambda = \infty$ on exige l'existence, pour K quelconque, d'une suite infinie de matrices A_m et de vecteurs $\mathbf{x}_m \neq 0$ tels que

$$\| A_m \mathbf{x}_m \|_{(m)} \geqslant K \| \mathbf{x}_m \|_{(m)}.$$

La quantité $\rho_m = \sup_{z \in S_m} |z|$ est le *rayon spectral de la matrice* et $\rho = \rho(\{A_m\}) = \sup_{z \in S(\{A_m\})} |z|$ le rayon spectral de la famille $\{A_m\}$. Evidemment $\rho_m \leqslant \| A_m \|_{(m)}$.

PROBLÈME. Démontrer que

$$S(\{A_m\}) = \overline{\bigcup_m S_m} \tag{3}$$

si $A_m = A_m^*$ et $\| \mathbf{x} \|_{(m)} = \| \mathbf{x} \|_3$ quels que soient m.

La proposition que nous venons de formuler entraîne que le spectre d'une famille de matrices symétriques de norme $\| \mathbf{x} \|_3$ est déterminé par les spectres des matrices de la famille.

La relation (3) n'est pas obligatoire pour des matrices non symétriques. Prenons pour modèle une famille simple de matrices non symétriques

$$A_m = \begin{pmatrix} \alpha & \beta & 0 & \dots & 0 \\ 0 & \alpha & \beta & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & \beta \\ 0 & 0 & 0 & \dots & \alpha \end{pmatrix}, \quad \alpha, \beta \neq 0.$$

Pour n'importe quel m l'ensemble S_m est composé du point $z = \alpha$ et l'ensemble $\overline{\bigcup_m S_m}$ coïncide donc avec le point α. Montrons que l'ensemble S contient également tous les points du cercle $|z - \alpha| \leqslant |\beta|$. Soit w un complexe quelconque tel que $|w| \leqslant 1$ et soit $\mathbf{x}_m$ un vecteur de composantes

$$(\mathbf{x}_m)_j = w^{j-1}\left(1 - \frac{j-1}{m}\right).$$

On a les égalités suivantes:

$$(A_m \mathbf{x}_m)_j = \alpha (\mathbf{x}_m)_j + \beta (\mathbf{x}_m)_{j+1} =$$

$$= \alpha w^{j-1}\left(1 - \frac{j-1}{m}\right) + \beta w^j \left(1 - \frac{j}{m}\right) =$$

$$= (\alpha + \beta w) w^{j-1}\left(1 - \frac{j-1}{m}\right) - \beta \frac{w^j}{m} \quad \text{pour } j < m,$$

$$(A_m \mathbf{x}_m)_m = \alpha (\mathbf{x}_m)_m = \alpha w^{m-1}\left(1 - \frac{m-1}{m}\right) =$$

$$= (\alpha + \beta w) w^{m-1}\left(1 - \frac{m-1}{m}\right) - \beta \frac{w^m}{m}.$$

Ainsi,

$$(A_m \mathbf{x}_m)_j = (\alpha + \beta w)(\mathbf{x}_m)_j - \beta \frac{w^j}{m} \text{ quel que soit } j.$$

D'où

$$\| A_m \mathbf{x}_m - (\alpha + \beta w)\, \mathbf{x}_m \|_3 = \sqrt{\sum_{j=1}^{m} \left| \beta \frac{w^j}{m} \right|^2} \leqslant \sqrt{m \left(\frac{\beta}{m} \right)^2} = \frac{|\beta|}{\sqrt{m}}.$$

Par ailleurs

$$\| \mathbf{x}_m \|_3 = \sqrt{\sum_{j=1}^{m} \left| w^{j-1} \left(1 - \frac{j-1}{m} \right) \right|^2} \geqslant 1.$$

Donc,

$$\| A_m \mathbf{x}_m - (\alpha + \beta w)\, \mathbf{x}_m \|_3 \leqslant \left(\frac{|\beta|}{\sqrt{m}} \right) \| \mathbf{x}_m \|_3.$$

Comme pour un $\varepsilon > 0$ il existe un m tel que $\frac{|\beta|}{\sqrt{m}} < \varepsilon$, l'inégalité obtenue veut dire que le point $\alpha + \beta w$ appartient au spectre de la famille de matrices $\{A_m\}$. Puisque tout point du cercle $|z - \alpha| \leqslant |\beta|$ s'écrit sous cette forme avec $|w| \leqslant 1$, nous avons démontré l'appartenance de ce cercle au spectre S.

Posons $\varphi_n = \sup_m \| (A_m)^n \|_{(m)}$.

THÉORÈME 1. *Si* $\rho(\{A_m\}) = \rho < \infty$, *on a pour tout* $\sigma > \rho$ *l'estimation*

$$\varphi_n \leqslant k(\sigma)\, \sigma^n \tag{4}$$

avec n quelconque; ici $k(\sigma)$ *ne dépend pas de n.*

DÉMONSTRATION. Soit D une matrice arbitraire de norme $\| D \| < d < 1$. Considérons l'équation matricielle

$$X = DX + C. \tag{5}$$

On a évidemment

$$\| X \| \leqslant \| D \| \, \| X \| + \| C \| \leqslant d \| X \| + \| C \|,$$

d'où

$$\| X \| \leqslant (1 - d)^{-1} \| C \|. \tag{6}$$

Le système (5) est un système d'équations linéaires en éléments de la matrice X. Puisque $X = 0$ pour $C = 0$ (vu (6)), ce système admet une solution unique pour toute matrice C. Si $C = E$, on a $X = (E - D)^{-1}$, et, selon (6),

$$\| (E - D)^{-1} \| \leqslant (1 - d)^{-1}. \tag{7}$$

On a, pour tout n,

$$E - D^n = (E - D)(E + D + \ldots + D^{n-1}),$$

et, après multiplication par $(E - D)^{-1}$,

$$(E - D)^{-1} - (E - D)^{-1} D^n = E + D + \ldots + D^{n-1};$$

il s'ensuit

$$(E - D)^{-1} - (E + \ldots + D^{n-1})\| \leqslant$$
$$\leqslant \| (E - D)^{-1}\| \| D^n\| \leqslant (1 - d)^{-1} d^n. \quad (8)$$

Soient B une matrice quelconque et β un nombre supérieur à sa norme: $\| B\| < \beta$; si $|z| \geqslant \beta$, on a $\| z^{-1}B\| \leqslant \frac{\|B\|}{\beta} = d < 1$. Faisant $D = z^{-1}B$ dans (8) il vient

$$\left\| (E - z^{-1}B)^{-1} - \sum_{j=0}^{n-1} (z^{-1}B)^j \right\| \leqslant \left(1 - \frac{\|B\|}{\beta}\right)^{-1} \left(\frac{\|B\|}{\beta}\right)^n.$$

Ainsi, la série matricielle $\sum_{j=0}^{\infty} z^{-j}B^j$ converge uniformément vers la matrice $(E - z^{-1}B)^{-1}$ dans le domaine $|z| \geqslant \beta$. La formule des coefficients de la série de Laurent nous permet d'écrire

$$B^n = \frac{1}{2\pi i} \oint_{|z|=\beta} z^{n-1} (E - z^{-1}B)^{-1} dz,$$

ou encore

$$B^n = \frac{1}{2\pi i} \oint_{|z|=\beta} z^n (zE - B)^{-1} dz. \quad (9)$$

Les éléments de la matrice $(zE - B)^{-1}$ sont égaux au signe près aux rapports des mineurs de la matrice $zE - B$ à son déterminant. Ils sont donc analytiques partout sauf aux points z, zéros de $\det(zE - B)$, i. e. en dehors du spectre de B. Par conséquent, dans (9) on peut passer à l'intégration le long de tout contour renfermant le spectre de B. Soit $\sigma > \rho$ quelconque. Les spectres S_m de toutes les matrices A_m se trouvent à l'intérieur du cercle $|z| \leqslant \sigma$. Aussi

$$A_m^n = \frac{1}{2\pi i} \oint_{|z|=\sigma} z^n (zE - A_m)^{-1} dz. \quad (10)$$

En démontrant le théorème 1 on s'adressera au

LEMME. *Si* $\sigma > \rho$, *alors*

$$k(\sigma) = \sup_{|z|=\sigma,\, m} \| (zE - A_m)^{-1} \|_{(m)} < \infty.$$

DÉMONSTRATION. Puisque $\sigma > \rho$, on a, pour $|z| = \sigma$ et tout m, $\det(zE - A_m) \neq 0$ et $\|(zE - A_m)^{-1}\|_{(m)} < \infty$.

Supposons que le lemme n'est pas juste. Il existe alors une suite $(z(q), m(q))$ telle que

$$|z(q)| = \sigma, \quad w(q) = \|(z(q)E - A_{(m(q))})^{-1}\|_{(m(q))} \to \infty$$

pour q tendant vers l'infini. Soit z_0, $|z_0| = \sigma$, la valeur d'adhérence de la suite $z(q)$. Afin d'éviter tout changement de notations supplémentaire, supposons que $z(q) \to z_0$. Par définition de la norme d'une matrice, quel que soit q il existe un vecteur $\mathbf{x}_{m(q)}$ tel que

$$\|(z(q)E - A_{m(q)})^{-1}\mathbf{x}_{m(q)}\|_{(m(q))} \geqslant \frac{w(q)}{2}\|\mathbf{x}_{m(q)}\|_{(m(q))},$$

ou, ce qui revient au même,

$$\|(z(q)E - A_{m(q)})\mathbf{y}_{m(q)}\|_{(m(q))} \leqslant \frac{2}{w(q)}\|\mathbf{y}_{m(q)}\|_{(m(q))}$$

avec

$$\mathbf{y}_{m(q)} = (z(q)E - A_{m(q)})^{-1}\mathbf{x}_{m(q)}.$$

Il en découle l'inégalité

$$\|A_{m(q)}\mathbf{y}_{m(q)} - z_0\mathbf{y}_{m(q)}\|_{(m(q))} \leqslant \left(|z(q) - z_0| + \frac{2}{w(q)}\right)\|\mathbf{y}_{m(q)}\|_{(m(q))}.$$

Puisque $w(q) \to \infty$ par hypothèse et $|z(q) - z_0| \to 0$, il en résulte que le point z_0, $|z_0| = \sigma$, est un point du spectre: contradiction. Le lemme est démontré.

En évaluant la norme du second membre de (10), il vient

$$\|A_m^n\|_{(m)} \leqslant \frac{1}{2\pi}\oint_{|z|=\sigma} \sigma^n k(\sigma)\,|dz| = k(\sigma)\,\sigma^n.$$

D'où l'inégalité (4).

THÉORÈME 2. *Si* $\rho > 1$, *alors*

$$\varphi = \sup_n \varphi_n = \sup_{n,\,m} \|A_m^n\|_{(m)} = \infty.$$

DÉMONSTRATION. Si, pour une famille donnée de matrices, on a $\varphi_1 = \infty$, elle vérifie le théorème.

Considérons le cas $\varphi_1 < \infty$. Pour démontrer le théorème 2 il suffit de prouver que, quel que soit K, il existe m et n tels que $\|A_m^n\|_{(m)} \geqslant K$.

Si $\rho > 1$, il existe un $z \in S$, $|z| > 1$. Prenons K arbitraire et choisissons n à partir de la condition $|z|^n \geqslant 2K$. Posons $\delta = K/(nb^{n-1})$, où $b = \max(\varphi_1, |z|)$. z étant un point du spectre, il

existe m et un vecteur $\mathbf{x}_m$ vérifiant

$$\| A_m\mathbf{x}_m - z\mathbf{x}_m \|_{(m)} \leqslant \delta \| \mathbf{x}_m \|_{(m)}.$$

On a évidemment

$$\| A_m^{n-i-1} z^i (A_m\mathbf{x}_m - z\mathbf{x}_m) \|_{(m)} \leqslant \| A_m \|_{(m)}^{n-i-1} |z|^i \delta \| \mathbf{x}_m \|_{(m)} \leqslant$$
$$\leqslant \varphi_1^{n-i-1} |z|^i \delta \| \mathbf{x}_m \|_{(m)} \leqslant b^{n-1}\delta \| \mathbf{x}_m \|_{(m)} = \frac{K}{n} \| \mathbf{x}_m \|_{(m)}. \quad (11)$$

On a l'identité

$$A_m^n\mathbf{x}_m - z^n\mathbf{x}_m = \sum_{i=0}^{n-1} A_m^{n-i-1} z^i (A_m\mathbf{x}_m - z\mathbf{x}_m).$$

Après avoir évalué la norme de chaque terme à droite à l'aide de (11), il vient

$$\| A_m^n\mathbf{x}_m - z^n\mathbf{x}_m \|_{(m)} \leqslant K \| \mathbf{x}_m \|_{(m)},$$

d'où

$$\| A_m^n\mathbf{x}_m \|_{(m)} \geqslant |z|^n \| \mathbf{x}_m \|_{(m)} - K \| \mathbf{x}_m \|_{(m)} =$$
$$= (|z|^n - K) \| \mathbf{x}_m \|_{(m)} \geqslant K \| \mathbf{x}_m \|_{(m)}.$$

Nous avons trouvé les m et n pour lesquels $\| A_m^n \|_{(m)} \geqslant K$, ce qui démontre le théorème.

Les affirmations des théorèmes entraînent que le rayon spectral d'une famille d'opérateurs détermine pour beaucoup les propriétés du processus itératif. En effet, si $\rho < 1$, alors, d'après le théorème 1, les normes $\| A_m^n \|_{(m)}$ sont uniformément bornées, et le processus itératif doit être insensible aux arrondis; si, par contre, $\rho > 1$, on a $\sup_{n,\,m} \| A_m^n \|_{(m)} = \infty$ conformément au théorème 2, et, pour certaines des matrices A_m, le processus $\mathbf{x}^{n+1} = A_m\mathbf{x}^n + \mathbf{b}$ ou bien divergera, ou bien sera assez sensible aux arrondis.

Les théorèmes démontrés n'autorisent aucun jugement sur la stabilité d'un processus itératif concret par rapport aux arrondis. Il y a plus. En l'absence de relations nettes entre les éléments de la matrice A on ne sait pas exactement à quelle famille convient-il de la rapporter. Et quand bien même le problème du choix de $\{A_m\}$ et de la recherche de son spectre serait résolu, la question de la stabilité du processus itératif reste entière. Pour un nombre réel de caractères binaires et une précision fixe du calcul, il se peut que les normes $\| A_m^n \|_{(m)}$ croissant lentement avec n et m soient à préférer aux mêmes normes uniformément bornées par une constante importante. Aucune conclusion pratique concrète n'est donc possible. Néanmoins, nous avons établi que les matrices non symétriques de dimension élevée jouissaient de propriétés qui nous étaient inconnues. Elles possèdent, en particulier, à côté de valeurs et vecteurs

propres, des valeurs et vecteurs « presque propres », c'est-à-dire des λ et y tels que $By \approx \lambda y$. La dernière égalité implique $B^n y \approx \approx \lambda^n y$, donc la norme d'une erreur proportionnelle à y augmente d'environ $|\lambda|^n$ fois au bout de n pas d'itération. La présence de ces valeurs « presque propres » est justement la cause des effets déplorables dont nous avons parlé au début du paragraphe précédent.

§ 6. Procédé δ^2 d'évaluation pratique de l'erreur et d'accélération de la convergence

Voyons comment on évalue l'erreur sur une solution approchée d'un système d'équations. Si $\mathbf{X}^*$ est une solution approchée du système $A\mathbf{X} = \mathbf{b}$, on écrit $A(\mathbf{X}^* - \mathbf{X}) = A\mathbf{X}^* - \mathbf{b}$. D'où l'estimation d'erreur

$$\|\mathbf{X}^* - \mathbf{X}\| = \|A^{-1}(A\mathbf{X}^* - \mathbf{b})\| \leqslant \|A^{-1}\| \|A\mathbf{X}^* - \mathbf{b}\|;$$

elle n'est guère applicable vu la complexité de l'évaluation de $\|A^{-1}\|$. Aussi, en analysant pratiquement l'erreur d'approximations dues aux méthodes itératives, recourt-on plutôt à une autre évaluation peu rigoureuse, mais plus simple.

Rappelons le critère de rationalité d'une évaluation pratique d'erreur: on dit que $\mathbf{v}^n$ est l'erreur réelle sur l'approximation $\mathbf{x}^n$, qui tend vers X lorsque $n \to \infty$, si

$$\|\mathbf{V}^n - (\mathbf{x}^n - \mathbf{X})\| / \|\mathbf{x}^n - \mathbf{X}\| \to 0 \text{ pour } n \to \infty. \tag{1}$$

Evidemment, on a alors $\|\mathbf{v}^n\| \sim \|\mathbf{x}^n - \mathbf{X}\|$.

Considérons la méthode itérative simple $\mathbf{x}^{n+1} = B\mathbf{x}^n + \mathbf{c}$; bornons-nous, afin d'être simple au souhait, au cas où B est de structure simple (sa forme de Jordan est diagonale et B possède donc un système complet de vecteurs propres).

Soient λ_i, $i = 1, \ldots, m$, les valeurs propres de B qui sont numérotées dans l'ordre de décroissance de $|\lambda_i|$, de plus,

$$1 > |\lambda_1| > |\lambda_2| \geqslant |\lambda_3| \geqslant \ldots \geqslant |\lambda_m|,$$

et $\mathbf{e}_i$, $\|\mathbf{e}_i\| = 1$ les vecteurs propres associés du système complet.

Développons le vecteur $\mathbf{r}^0$ suivant la base $\mathbf{e}_i$: $\mathbf{r}^0 = \sum c_i \mathbf{e}_i$. Alors

$$\mathbf{r}^n = \mathbf{x}^n - \mathbf{X} = B^n \mathbf{r}^0 = \sum c_i \lambda_i^n \mathbf{e}_i = c_1 \lambda_1^n \mathbf{e}_1 + O(|\lambda_2|^n). \tag{2}$$

Ici et dans la suite $x_n = y_n + O(\varepsilon_n)$ a le sens suivant:

$$\|x_n - y_n\| = O(\varepsilon_n) \quad \text{pour} \quad n \to \infty.$$

Indiquons un procédé d'approximation du vecteur $\mathbf{w}^n = c_1 \lambda_1^n \mathbf{e}_1$ qui tienne compte de l'information obtenue en cours du calcul. On a,

selon (2),

$$x^{n-2}-X=w^n\lambda_1^{-2}+O(|\lambda_2|^n),$$
$$x^{n-1}-X=w^n\lambda_1^{-1}+O(|\lambda_2|^n),$$
$$x^n-X=w^n+O(|\lambda_2|^n).$$

En retranchant membre à membre les expressions voisines, il vient

$$x^{n-1}-x^{n-2}=w^n(1-\lambda_1^{-1})\lambda_1^{-1}+O(|\lambda_2|^n),$$
$$x^n-x^{n-1}=w^n(1-\lambda_1^{-1})+O(|\lambda_2|^n). \tag{3}$$

D'où

$$(x^n-x^{n-1},\ x-x^{n-1})=\|w^n\|^2|1-\lambda_1^{-1}|^2+O(\|w^n\||\lambda_2|^n),$$
$$(x^{n-1}-x^{n-2},\ x^n-x^{n-1})=\|w^n\|^2|1-\lambda_1^{-1}|^2\lambda_1^{-1}+O(\|w^n\||\lambda_2|^n). \tag{4}$$

Posons

$$\lambda_1^{(n)}=\frac{(x^n-x^{n-1},\ x^n-x^{n-1})}{(x^{n-1}-x^{n-2},\ x^n-x^{n-1})}.$$

Prenons les égalités (4) et sous l'hypothèse $c_1\neq 0$ divisons le numérateur et le dénominateur par $\|w^n\|^2|1-\lambda_1^{-1}|^2\lambda_1^{-1}$:

$$\lambda_1^{(n)}=\frac{\lambda_1+O\left(\frac{|\lambda_2|^n}{\|w^n\|}\right)}{1+O\left(\frac{|\lambda_2|^n}{\|w^n\|}\right)}.$$

Puisque

$$\|w^n\|=|c_1||\lambda_1|^n, \tag{5}$$

on a

$$\lambda^{(n)}=\lambda_1+O\left(\left|\frac{\lambda_2}{\lambda_1}\right|^n\right). \tag{6}$$

Divisant la seconde relation (3) par $1-(\lambda_1^{(n)})^{-1}$, on trouve

$$\frac{x^n-x^{n-1}}{1-(\lambda_1^{(n)})^{-1}}=w^n\frac{1-\lambda_1^{-1}}{1-(\lambda_1^{(n)})^{-1}}+O(|\lambda_2|^n)=w^n+w^n\frac{\lambda_1-\lambda_1^{(n)}}{\lambda_1(\lambda_1^{(n)}-1)}+O(|\lambda_2|^n).$$

(5), (6) entraînent $\|w^n(\lambda_1-\lambda_1^{(n)})\|=O(|\lambda_2|^n)$, et donc

$$\frac{x^n-x^{n-1}}{1-(\lambda_1^{(n)})^{-1}}=w^n+O(|\lambda_2|^n).$$

Il s'ensuit, compte tenu de (2),

$$x^n-X=v^n+O(|\lambda_2|^n)$$

avec $v^n=\frac{x^n-x^{n-1}}{1-(\lambda_1^{(n)})^{-1}}$; notons que, conformément à (3), (6), $\|v^n\|=$ $=|c_1||\lambda_1|^n+O(|\lambda_2|^n)$. Il en découle que v^n vérifie (1), ce qui autorise à le prendre pour erreur réelle sur l'approximation x^n.

Pour $c_1 = \ldots = c_l = 0$, $c_{l+1} \neq 0$, nos raisonnements restent valables si $|\lambda_{l+1}| > |\lambda_{l+2}|$; on remplace dans toutes les relations λ_i, c_i, e_i $(i = 1, 2)$ par λ_{l+i}, c_{l+i}, e_{l+i}. Ainsi, nous avons évalué la solution approchée par le *procédé* δ^2.

Si on pose $y^n = x^n - v^n$, alors $y^n - X = O(|\lambda_2|^n)$, et y^n est en général une meilleure condition initiale des itérations devant x^n. En affinant de temps en temps les résultats on arrive parfois à réduire sensiblement le nombre d'itérations.

Pour que soit vraie l'égalité approchée

$$x^n - X \approx v^n, \tag{7}$$

il est nécessaire que le terme $c_1\lambda_1^n e_1$ à droite dans $x^n - X = \sum_i c_i\lambda_i^n e_i$ soit prépondérant. S'il est bien ainsi, les vecteurs $x^n - x^{n-1}$, $x^{n-1} - x^{n-2}$ sont approximativement proportionnels et

$$\mu_n = \frac{|(x^{n-1} - x^{n-2},\ x^{n-1} - x^n)|}{\|x^{n-1} - x^{n-2}\| \, \|x^{n-1} - x^n\|} \approx 1.$$

La condition $\mu_n \approx 1$ est donc nécessaire pour que soient vraies nos constructions ci-dessus. On peut donc la prendre pour critère d'applicabilité pratique de (7).

Le schéma suivant de la méthode itérative simple avec recours au procédé δ^2 d'accélération de la convergence paraît par exemple raisonnable. On se donne un certain $\eta > 0$ petit ; si, à une itération, $\mu_n \geqslant 1 - \eta$, on calcule le vecteur v^n et les itérations suivantes démarrent à partir de l'approximation initiale y^n. Le processus s'arrête si $\mu_n \geqslant 1 - \eta$ et $\|v^n\| \leqslant \varepsilon$, où ε est la précision voulue.

Pour η très faible, la condition $\mu_n \geqslant 1 - \eta$ n'est satisfaite qu'après de nombreuses itérations et la convergence n'est pas améliorée. Pour η grand, les relations à la base de nos constructions ne sont plus strictes et il se peut que le procédé δ^2 ralentisse le processus itératif. La situation est compliquée par la présence d'arrondis, et le schéma décrit appelle une mise au point sur un nombre important d'exemples afin de pouvoir choisir un η optimal et déterminer la borne inférieure des valeurs de ε permettant l'algorithme en question. Si un processus itératif homogène se trouve remanié (dans notre cas, il l'a été lors du passage de x^n à y^n), il serait bon de vérifier si une telle modification n'en détériore pas la qualité. On prend comme critère d'opportunité du remaniement une certaine relation entre les normes des résidus pour x^n et y^n, par exemple une inégalité de la forme

$$\|(E - B)\, y^n - c\| \leqslant q \, \|(E - B)\, x^n - c\|.$$

Quand nous avons parlé de la nécessité d'indiquer la borne inférieure des valeurs de ε, nous nous sommes guidé sur la circonstance suivante. Soit, pour fixer les idées, $\lambda_1 > 0$. Le calcul de x^n d'après x^{n-1} donné suffit déjà pour que les arrondis perturbent le résultat d'une quantité δx^n dotée d'une norme de l'ordre de ρ. Cela risque d'entraîner une perturbation δv^n de norme en $(1 - \lambda_1)^{-1}\rho$. Il en résulte que pour $\varepsilon < (1 - \lambda_1)^{-1}\rho$ le processus itératif est susceptible de ne jamais finir.

Nos raisonnements prouvent que la mise en pratique d'une méthode engendre des situations qui exigent une préparation mathématique sérieuse et des expériences numériques en quantité considérable. Aussi, toute « simple » que paraisse la technique itérative simple, l'élaboration d'un programme standard est-elle complètement justifiée.

§ 7. Optimisation de la vitesse de convergence des processus itératifs

Considérons un procédé très simple de résolution du système $Ax = b$ par itérations:

$$x^{n+1} = x^n - \alpha (Ax^n - b).$$

On sait déjà que sa vitesse de convergence dépend essentiellement du module maximal des valeurs propres de la matrice $B = E - \alpha A$. Si $\lambda_1, \ldots, \lambda_n$ sont les valeurs propres de A, on a

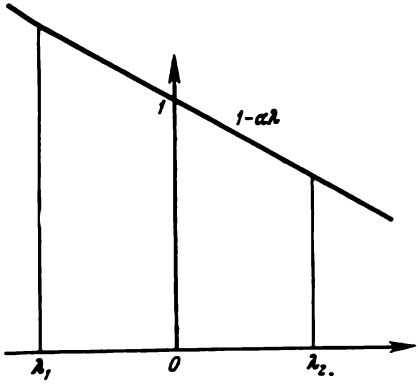


Fig. 6.7.1

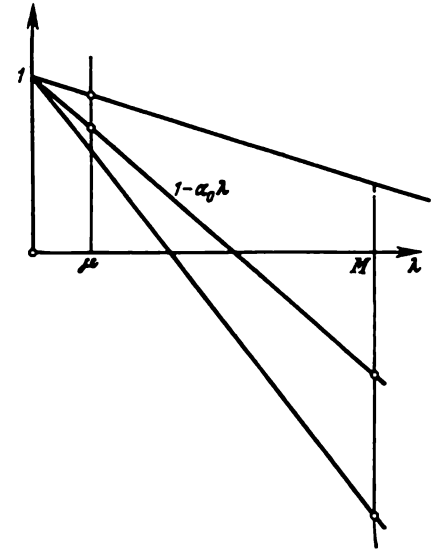


Fig. 6.7.2

$\max_i |\lambda_i(B)| = \max_i |1 - \alpha\lambda_i|$. La figure 6.7.1 montre que s'il y a des valeurs propres réelles de signes contraires, ce maximum est supérieur à l'unité et le processus itératif diverge.

Voici un cas particulier avec toutes les $\lambda_i > 0$. Les valeurs λ_i sont connues fort rarement encore qu'un cas assez typique soit celui où on en connaît l'estimation:

$$0 < \mu \leqslant \lambda_i \leqslant M < \infty \text{ pour tous les } i.$$

La vitesse de convergence se caractérise par

$$\rho(\alpha) = \max_{\mu \leqslant \lambda \leqslant M} |1 - \alpha\lambda|.$$

Considérons le problème de minimisation de $\rho(\alpha)$ par le choix de α.

Pour trouver $\min_{\alpha} \rho(\alpha)$ on se servira de considérations géométriques (fig. 6.7.2). Il est clair que $\rho(\alpha) \geqslant 1$ pour $\alpha \leqslant 0$. Si $0 < \alpha \leqslant M^{-1}$, la fonction $1 - \alpha\lambda$ est non négative et décroît de façon monotone sur $[\mu, M]$, donc $\rho(\alpha) = 1 - \alpha\mu$. Pour $M^{-1} < \alpha$, la quantité $1 - \alpha M$ devient négative et son module augmente avec α ; pour un certain $\alpha = \alpha_0$, $1 - \alpha_0\mu$ devient égal à $-(1 - \alpha_0 M)$ et alors $\rho(\alpha_0) = |1 - \alpha_0\mu|$. Si $\alpha < \alpha_0$, alors $\rho(\alpha) = 1 - \alpha\mu > 1 - \alpha_0\mu = \rho(\alpha_0)$; si $\alpha_0 < \alpha$, on a $\rho(\alpha) \geqslant |1 - \alpha M| = M\alpha - 1 \geqslant M\alpha_0 - 1 = \rho(\alpha_0)$. La valeur $\alpha = \alpha_0$ est donc la valeur cherchée. En résolvant l'équation par rapport à α_0, on trouve $\alpha_0 = \frac{2}{M+\mu}$. D'où

$$\rho(\alpha_0) = \frac{M-\mu}{M+\mu}.$$

Nous avons optimisé un processus itératif particulier en nous appuyant sur une image concrète.

Sur l'exemple de systèmes de matrice $A > 0$ (ici et dans ce qui suit l'inégalité $A > 0$ signifie que A est une matrice autoconjuguée définie positive) examinons d'autres façons de poser les problèmes d'optimisation de la vitesse de convergence de processus itératifs. Une partie des résultats pourront jouer pour les matrices non symétriques.

Tout système $A\mathbf{x} = \mathbf{b}$ avec $\det A \neq 0$ est généralement réductible à une forme correspondant à $A > 0$ par une multiplication par la matrice A^* (on dit alors qu'il est *symétrisé*). En effet, le système $A^*A\mathbf{x} = A^*\mathbf{b}$ est équivalent au système initial, la matrice A^*A est autoconjuguée $(A^*A = (A^*A)^*)$ et définie positive $((A^*A\mathbf{x}, \mathbf{x}) = \|A\mathbf{x}\|^2 > 0$ pour $\mathbf{x} \neq 0)$. La symétrisation est à éviter car nous le verrons à la fin du paragraphe, elle conduit souvent à une détérioration des propriétés de la matrice.

Soit $\mathbf{x}^0$ l'approximation initiale et admettons que chaque approximation suivante est construite selon

$$\mathbf{x}^k = \mathbf{x}^0 - Q_{k-1}(A)(A\mathbf{x}^0 - \mathbf{b}), \tag{1}$$

$Q_{k-1}(A)$ étant un polynôme de degré $k - 1$ de la matrice A.

On sait du spectre de A qu'il appartient au segment $[\mu, M]$, $0 < \mu \leqslant M < \infty$. La solution $\mathbf{X}$ du système de départ vérifie

$$\mathbf{X} = \mathbf{X} - Q_{k-1}(A)(A\mathbf{X} - \mathbf{b}).$$

En retranchant cette égalité de (1), on obtient une équation donnant l'erreur

$$\mathbf{r}^k = \mathbf{r}^0 - Q_{k-1}(A)A\mathbf{r}^0 = P_k(A)\mathbf{r}^0$$

avec $P_k(\lambda) = 1 - \lambda Q_{k-1}(\lambda)$. La matrice $P_k(A)$ est symétrique en tant que fonction d'une matrice symétrique ; c'est pourquoi

$$\sup_{r^0 \neq 0} \frac{\| r^k \|_3}{\| r^0 \|_3} = \| P_k(A) \|_3 = \max_{\lambda \in S(A)} | P_k(\lambda) | ; \tag{2}$$

introduisons la notation $\sigma(g) = \max\limits_{\mu \leqslant \lambda \leqslant M} | g(\lambda) |$ et posons

$$\sigma_k = \inf_{Q_{k-1}} \sigma(P_k).$$

LEMME.

$$\sigma_k = \max_{\mu \leqslant \lambda \leqslant M} | 1 - \lambda Q^0_{k-1}(\lambda) | = \frac{1}{| t_k |} ; \tag{3}$$

ici

$$1 - \lambda Q^0_{k-1}(\lambda) = P^0_k(\lambda) = T_k\left(\frac{2\lambda - (M+\mu)}{M-\mu}\right)\left(T_k\left(-\frac{M+\mu}{M-\mu}\right)\right)^{-1}, \tag{4}$$

comme d'habitude T_k sont les polynômes de Tchebycheff, $t_k = T_k\left(-\frac{M+\mu}{M-\mu}\right)$.

DÉMONSTRATION. En utilisant $\max\limits_{|x| \leqslant 1} | T_k(x) | = 1$, on obtient $\sigma(P^0_k) = | t_k |^{-1}$. Supposons qu'il existe un polynôme $P_k(\lambda) = 1 - \lambda Q_{k-1}(\lambda)$ tel que

$$\sigma(P_k) < \sigma(P^0_k) = | t_k |^{-1}. \tag{5}$$

Considérons le polynôme $S_k(\lambda) = P^0_k(\lambda) - P_k(\lambda)$. Soit

$$\lambda^j = \frac{M+\mu}{2} + \frac{M-\mu}{2} \cos\frac{\pi j}{k}, \quad j = 0, \ldots, k.$$

Par suite de l'égalité $T_k(x) = \cos(k \arccos x)$, il vient

$$P^0_k(\lambda^j) = (-1)^j t_k^{-1}.$$

Puisque $\lambda^j \in [\mu, M]$, on a, conformément à (5),

$$| P_k(\lambda^j) | < | t_k |^{-1},$$

d'où

$$\operatorname{sg} S_k(\lambda^j) = \operatorname{sg} P^0_k(\lambda^j).$$

Les points $\lambda^0, \ldots, \lambda^k$ sont monotones sur le segment $[\mu, M]$; vu que $S_k(\lambda)$ change de signe en passant de chacun de ces points au suivant, il possède sur $[\mu, M]$ k zéros. De plus,

$$S_k(0) = P^0_k(0) - P_k(0) = 1 - 1 = 0.$$

Nous venons d'obtenir un polynôme de degré k qui a $k+1$ zéros ; par conséquent, $S_k(\lambda) \equiv 0$, $P_k(\lambda) \equiv P_k^0(\lambda)$ et

$$\max_{[\mu, M]} |P_k(\lambda)| = |t_k|^{-1},$$

ce qui est contraire à (5). Le lemme est démontré.

EXERCICE. Utiliser la formule

$$T_k(x) = \frac{(x+\sqrt{x^2-1})^k + (x-\sqrt{x^2-1})^k}{2}$$

pour vérifier la validité de l'égalité

$$t_k = \frac{(-1)^k}{2}(\lambda_0^k + \lambda_0^{-k}),$$

$$\lambda_0 = \frac{\sqrt{M}+\sqrt{\mu}}{\sqrt{M}-\sqrt{\mu}}.$$

De cette façon de représenter t_k il découle en particulier que

$$\sigma_k = \frac{1}{|t_k|} = \frac{1}{\operatorname{ch}(k \ln \lambda_0)} < 1 \quad \text{pour } k > 0.$$

Les constructions faites supposent plusieurs types de processus itératifs.

Dans un des cas on se donne une suite de valeurs $n_0 = 0 < < n_1 < n_2 \ldots$, et les approximations $\mathbf{x}^{n_i}$ sont définies par la formule de récurrence

$$\mathbf{x}^{n_{i+1}} = \mathbf{x}^{n_i} - Q_{q_i-1}^0(A)(A\mathbf{x}^{n_i} - \mathbf{b}) \tag{6}$$

avec $q_i = n_{i+1} - n_i$. On a les relations évidentes

$$\mathbf{r}^{n_{i+1}} = P_{q_i}^0(A)\,\mathbf{r}^{n_i},$$

$$\|\mathbf{r}^{n_{i+1}}\|_3 \leqslant \sigma_{q_i} \|\mathbf{r}^{n_i}\|_3$$

et finalement

$$\|\mathbf{r}^{n_p}\|_3 \leqslant \sigma_{q_0} \ldots \sigma_{q_{p-1}} \|\mathbf{r}^0\|_3.$$

Soit le cas $q_i \equiv k$, i.e. $n_i = ik$. En notant $\mathbf{x}^{n_i} = \mathbf{y}^i$, on écrit le processus itératif (6) sous la forme

$$\mathbf{y}^{i+1} = \mathbf{y}^i - Q_{k-1}^0(A)(A\mathbf{y}^i - \mathbf{b})$$

avec pour estimation d'erreur

$$\|\mathbf{y}^i - \mathbf{X}\|_3 \leqslant (\sigma_k)^i \|\mathbf{y}^0 - \mathbf{X}\|_3.$$

Un tel processus est dit *linéaire optimal à pas liés*. Pour le sac particulier $k=1$, on a selon (4)

$$P_1^0(\lambda)=1-\lambda Q_0^0(\lambda)=\frac{\frac{2\lambda-M-\mu}{M-\mu}}{-\frac{M+\mu}{M-\mu}}=1-\lambda\frac{2}{M+\mu},$$

$$t_1=\frac{1}{-\frac{M+\mu}{M-\mu}},\quad \sigma_1=\frac{1}{|t_1|}=\frac{M-\mu}{M+\mu}.$$

Le processus analogue *à pas séparés* se représente donc comme suit:

$$y^{i+1}=y^i-\frac{2}{M+\mu}(Ay^i-b)$$

avec pour estimation d'erreur

$$\|y^i-X\|_3\leqslant\left(\frac{M-\mu}{M+\mu}\right)^i\|y^0-X\|_3. \tag{7}$$

On vérifie que les coefficients des polynômes $Q_{k-1}^0(A)$ croissent rapidement avec k; on conçoit que, pour k grand, l'algorithme de calcul de x^n qui fait recours à l'information sur les valeurs desdits coefficients est très sensible à l'erreur de calcul. Aussi pour trouver x^k suit-on des voies détournées.

Mettons (1) sous la forme

$$x^k=P_k(A)x^0-(P_k(A)-E)A^{-1}b$$

et considérons-la comme la transformation $x^0\to x^k$.

Soient $P^{(1)}(\lambda)$ et $P^{(2)}(\lambda)$ des polynômes arbitraires et $P(\lambda)=P^{(1)}(\lambda)P^{(2)}(\lambda)$; effectuons successivement les transformations correspondant à $P^{(1)}$ et $P^{(2)}$:

$$\begin{aligned}x^{(1)}&=P^{(1)}(A)x^0-(P^{(1)}(A)-E)A^{-1}b,\\ x^{(2)}&=P^{(2)}(A)x^{(1)}-(P^{(2)}(A)-E)A^{-1}b.\end{aligned} \tag{8}$$

On a

$$\begin{aligned}x^{(2)}=P^{(2)}(A)P^{(1)}(A)x^0-P^{(2)}(A)(P^{(1)}(A)-E)A^{-1}b-\\ -(P^{(2)}(A)-E)A^{-1}b=P^{(2)}(A)P^{(1)}(A)x^0-\\ -(P^{(2)}(A)P^{(1)}(A)-E)A^{-1}b.\end{aligned}$$

Ainsi, les transformations successives (8) équivalent à la transformation correspondant au polynôme $P(A)$. Il s'ensuit que lorsque $P(\lambda)=P^{(1)}(\lambda)\ldots P^{(k)}(\lambda)$, la réalisation successive des transformations correspondant à $P^{(i)}(\lambda)$, $i=1,\ldots,k$, conduit au même résultat que la transformation correspondant à $P(\lambda)$.

Pour calculer $x^k=x^0-Q_{k-1}^0(A)(Ax^0-b)$ procédons comme suit. Décomposons

$$P_k^0(\lambda)=1-\lambda Q_{k-1}^0(\lambda)$$

en produit de facteurs

$$P_k^0(\lambda) = (1-\alpha_1^k\lambda)\ \ldots\ (1-\alpha_k^k\lambda)$$

et effectuons l'une après l'autre les transformations qui leur correspondent. Si $P(\lambda) = 1-\alpha_i^k\lambda$, alors $Q(\lambda) = (1-P(\lambda))\,\lambda^{-1} = \alpha_i^k$ et elles s'écrivent

$$\mathbf{x}^{(i)} = \mathbf{x}^{(i-1)} - \alpha_i^k (A\mathbf{x}^{(i-1)} - \mathbf{b}), \quad i = 1, \ \ldots, \ k. \tag{9}$$

Il vient finalement $\mathbf{x}^{(k)}$ égal à $\mathbf{x}^k$. Les zéros des polynômes $P_k^0(\lambda)$ sont $\lambda_i = (\alpha_i^k)^{-1}$, ceux des $T_k\left(\frac{2\lambda-(M+\mu)}{M-\mu}\right)$

$$\lambda_i = \frac{M+\mu}{2} + \frac{M-\mu}{2}\cos\frac{\pi(2i-1)}{2k},$$

$$i = 1, \ \ldots, \ k.$$

D'où

$$\alpha_i^k = \frac{2}{M+\mu+(M-\mu)\cos\frac{\pi(2i-1)}{2k}}. \tag{10}$$

Lorsque k est grand, l'algorithme de calcul selon (9), (10) est lui aussi sensible aux arrondis. Retranchant de (9) $\mathbf{X} = \mathbf{X} - \alpha_i^k(A\mathbf{X} - \mathbf{b})$, on trouve l'équation donnant l'erreur: $\mathbf{r}^{(i)} = (E - \alpha_i^k A)\,\mathbf{r}^{(i-1)}$, qui s'écrit en arrondissant:

$$\mathbf{r}^{(i)} = (E - \alpha_i^k A)\,\mathbf{r}^{i-1} + \boldsymbol{\rho}^{i-1}.$$

Exprimant chaque $\mathbf{r}^{(i)}$ par les itérées précédentes, on a l'égalité

$$\mathbf{r}^{(k)} = \prod_{i=1}^{k} (E - \alpha_i^k A)\,\mathbf{r}^{(0)} + \sum_{i=0}^{k-1} \prod_{i+2\leqslant j\leqslant k} (E - \alpha_j^k A)\,\boldsymbol{\rho}^i.$$

Ici sur $\mathbf{r}^{(0)}$ agit l'opérateur $P_k^0(A)$ de norme $\sigma_k = \left|\frac{1}{t_k}\right| < 1$ tandis que les opérateurs sur $\boldsymbol{\rho}^i$ peuvent être munis de normes fort importantes.

Dans la pratique, on rend un algorithme insensible aux arrondis par le « brassage » des quantités α_i^k. Pour $k = 2^l$ l'algorithme correspondant est le suivant. En prenant successivement $j = 1, \ \ldots, \ l$, on construit la permutation « la plus désordonnée » des nombres, soit $1, \ \ldots, \ 2^j$. Lorsque $j = 1$, elle comprend 1, 2. Soit $(b_1^{j-1}, \ \ldots, \ b_{2^{j-1}}^{j-1})$ une permutation donnée; la suivante s'écrit alors $(b_1^{j-1}, \ 2^j + 1 - b_1^{j-1}, \ b_2^{j-1}, \ 2^j + 1 - b_2^{j-1}, \ \ldots)$. Pour $l = 4$ par exemple cette permutation a la forme (1, 16, 8, 9, 4, 13, 5, 12, 2, 15, 7, 10, 3, 14, 6, 11). En se donnant $k = 2^l$, construisons le tableau des nombres $b_1^l, \ \ldots, \ b_{2^l}^l$ et effectuons les itérations (9) pour

$$\alpha_i^k = \frac{2}{M+\mu+(M-\mu)\cos\frac{\pi(2b_i^l-1)}{2k}}.$$

Proposons-nous de construire un processus itératif qui donne l'approximation

$$\mathbf{y}^k = \mathbf{x}^0 - Q_{k-1}^0(A)(A\mathbf{x}^0 - \mathbf{b}) \tag{11}$$

au bout de k multiplications de la matrice A par un vecteur. On a conformément à (4)

$$Q_{k-1}^0(A) = \left(E - T_k\left(\frac{2A-(M+\mu)E}{M-\mu}\right)t_k^{-1}\right)A^{-1}.$$

Notons que $Q_{-1}^0(A) = (E-E)A^{-1} = 0$ et $\mathbf{y}^0 = \mathbf{x}^0$. Portant l'expression de $Q_{k-1}^0(A)$ dans (11) et multipliant les deux membres par t_k, il vient

$$t_k\mathbf{y}^k = T_k\left(\frac{2A-(M+\mu)E}{M-\mu}\right)(\mathbf{x}^0 - A^{-1}\mathbf{b}) + t_kA^{-1}\mathbf{b}.$$

La validité de cette relation lorsque $k=n-1$, $k=n$, $k=n+1$ entraîne

$$t_{n+1}\mathbf{y}^{n+1} - 2\frac{2A-(M+\mu)E}{M-\mu}t_n\mathbf{y}^n + t_{n-1}\mathbf{y}^{n-1} =$$

$$= \left\{T_{n+1}\left(\frac{2A-(M+\mu)E}{M-\mu}\right) - 2\frac{2A(M+\mu)E}{M-\mu}T_n\left(\frac{2A-(M+\mu)E}{M-\mu}\right) + \right.$$

$$\left. + T_{n-1}\left(\frac{2A-(M+\mu)E}{M-\mu}\right)\right\}(\mathbf{x}^0 - A^{-1}\mathbf{b}) +$$

$$+ \left(t_{n+1}E - 2\frac{2A-(M+\mu)E}{M-\mu}t_n + t_{n-1}E\right)A^{-1}\mathbf{b}. \tag{12}$$

Les polynômes de Tchebycheff sont liés par la relation de récurrence

$$T_{n+1}(x) - 2xT_n(x) + T_{n-1}(x) = 0. \tag{13}$$

Si l'on substitue à x une matrice, en particulier $\frac{2A-(M+\mu)E}{M-\mu}$, la relation reste stricte. Aussi la matrice entre accolades est-elle nulle. Par suite de (13) et de l'égalité $t_n = T_n\left(-\frac{M+\mu}{M-\mu}\right)$ on a

$$t_{n+1} - 2\left(-\frac{M+\mu}{M-\mu}\right)t_n + t_{n-1} = 0 \tag{14}$$

et la relation (12) se récrit

$$t_{n+1}\mathbf{y}^{n+1} - 2\left(-\frac{M+\mu}{M-\mu}\right)t_n\mathbf{y}^n + t_{n-1}\mathbf{y}^{n-1} = 4\frac{t_n}{M-\mu}(A\mathbf{y}^n - \mathbf{b}). \tag{15}$$

Nous avons ici la relation de récurrence cherchée. Mettons-la sous une forme qui soit d'un usage plus commode. Vu (14) l'égalité (15) s'écrit

$$t_{n+1}\mathbf{y}^{n+1} - (t_{n+1} + t_{n-1})\mathbf{y}^n + t_{n-1}\mathbf{y}^{n-1} = -\frac{2}{M+\mu}(t_{n+1} + t_{n-1})(A\mathbf{y}^n - \mathbf{b}),$$

ou encore

$$\mathbf{y}^{n+1} = \mathbf{y}^n + \omega_n\omega_{n-1}(\mathbf{y}^n - \mathbf{y}^{n-1}) - \frac{2}{M+\mu}(1+\omega_n\omega_{n-1})(A\mathbf{y}^n - \mathbf{b}) \quad (16)$$

avec

$$\omega_n = \left|-\frac{t_n}{t_{n+1}}\right.$$

La division par t_{n+1} des deux membres de (14) fournit

$$1 - \frac{2M+\mu}{M-\mu}\omega_n + \omega_n\omega_{n-1} = 0,$$

d'où

$$\omega_n = \frac{1}{2\dfrac{M+\mu}{M-\mu} - \omega_{n-1}}. \quad (17)$$

Ainsi, on peut calculer par récurrence ω_n à l'aide de (17), puis les vecteurs $\mathbf{y}^{n+1}$ moyennant (16). Pour obtenir les vecteurs $\mathbf{y}^1, \ldots$ $\ldots, \mathbf{y}^n$ il faut n multiplications de la matrice par le vecteur, $O(n)$ multiplications de vecteurs par les nombres, additions de vecteurs et opérations sur les nombres. Tous les k vérifient, par suite de (2)-(4), l'estimation

$$\| \mathbf{y}^k - \mathbf{X} \|_3 \leqslant | t_k |^{-1} \| \mathbf{x}^0 - \mathbf{X} \|_3. \quad (18)$$

Le processus itératif (16), (17) est dit *linéaire optimal*. En effet, on montre que toute approximation raisonnable de la solution telle que la matrice A ne participe qu'à k opérations de multiplication par le vecteur se met sous la forme (1). En réalisant le processus (16), (17) chaque ensemble de k applications de A fournit un résultat optimal au sens de (2)-(3). On voit donc que le processus (16), (17) converge plus rapidement que (6). Cependant, en raison du facteur occupation mémoire on lui préférera parfois (6).

Voici une évaluation plus concrète de la vitesse de convergence des processus itératifs construits.

D'après la formule explicite des polynômes de Tchebycheff, $T_n(x) = \frac{\lambda_1^n + \lambda_2^n}{2}$ avec $\lambda_{1,2} = x \pm \sqrt{x^2 - 1}$. Lorsque $x > 1$, on a $\lambda_1 = x + \sqrt{x^2 - 1} > 1$, $0 < \lambda_2 = \frac{1}{\lambda_1} < 1$. Aussi, pour n grands,

$$T_n(x) \sim \frac{\lambda_1^n}{2}, \quad (19)$$

$T_n(x) > \frac{\lambda_1^n}{2}$. Il en résulte $|t_n|^{-1} \leqslant 2\lambda_0^{-n}$, où $\lambda_0 = \frac{M+\mu}{M-\mu} + \sqrt{\left(\frac{M+\mu}{M-\mu}\right)^2 - 1}$, et, conformément à (18), pour le processus

itératif optimal

$$\| y^n - X \|_3 \leqslant 2\lambda_0^{-n} \| x^0 - X \|_3.$$

Lorsque $2\lambda_0^{-n} \leqslant \varepsilon$, la norme de l'erreur diminue d'au moins ε^{-1} fois. D'où l'évaluation du nombre d'itérations requis

$$n \geqslant n_1 = \log_{\lambda_0} \frac{2}{\varepsilon} = (\ln \lambda_0)^{-1} \ln \frac{2}{\varepsilon}.$$

On vérifie aisément que

$$\lambda_0 = \frac{\sqrt{M}+\sqrt{\mu}}{\sqrt{M}-\sqrt{\mu}} = \frac{1+\sqrt{\frac{\mu}{M}}}{1-\sqrt{\frac{\mu}{M}}}.$$

Dans beaucoup de problèmes M/μ est très important; pour $M/\mu \to \infty$, il vient

$$\lambda_0 = 1 + 2\sqrt{\frac{\mu}{M}} + O\left(\frac{\mu}{M}\right);$$

donc,

$$\ln \lambda_0 \sim 2\sqrt{\frac{\mu}{M}}, \quad n_1 \sim \frac{1}{2}\sqrt{\frac{M}{\mu}} \ln \frac{2}{\varepsilon}.$$

Faisons la comparaison avec le processus linéaire optimal à pas séparés dont l'erreur, selon (7), est évaluée par

$$\| x^n - X \|_3 \leqslant \left(\frac{M-\mu}{M+\mu}\right)^n \| x^0 - X \|_3;$$

l'estimation du nombre d'itérations s'écrit alors

$$n \geqslant n_2 = \left(\ln \frac{M+\mu}{M-\mu}\right)^{-1} \ln \frac{1}{\varepsilon}.$$

Si $\frac{M}{\mu} \to \infty$, alors

$$\ln \frac{M+\mu}{M-\mu} \sim 2\frac{\mu}{M} \quad \text{et} \quad n_2 \sim \frac{1}{2}\frac{M}{\mu} \ln \frac{1}{\varepsilon}.$$

Ainsi, le processus optimal fournit un gain de nombre d'opérations d'environ $\sqrt{\frac{M}{\mu}}$ fois.

PROBLÈMES. 1. Soit un processus itératif

$$x^i = x^{i-1} - \alpha_{i-1}(Ax^{i-1} - b), \quad i = 1, 2, \ldots.$$

Supposons p impair et les ensembles $\alpha_0, \ldots, \alpha_{p^k-1}$, k quelconque, coïncidant avec

$$\gamma_{i-1} = 2\left(M + \mu + (M-\mu)\cos\frac{\pi(2i-1)}{2p^k}\right)^{-1}, \quad i = 1, \ldots, p^k.$$

Vérifier que pour tous $i = p^k$, les approximations $\mathbf{x}^i$ coïncident avec les approximations du processus itératif linéaire optimal.

2. Soit

$$\alpha_0 = \frac{2}{M+\mu}, \quad \alpha_1 = \frac{1}{M}, \quad \alpha_2 = \frac{1}{\mu}$$

et les ensembles

$$\alpha_{2^k+1}, \ \ldots, \ \alpha_{2^{k+1}}$$

coïncident, quel que soit $k > 1$, avec les ensembles

$$\gamma_i = 2\left(M+\mu+(M-\mu)\cos\frac{\pi(2i-1)}{2^{k+1}}\right)^{-1}, \quad i = 1, \ \ldots, \ 2^k.$$

Montrer que pour ce processus itératif et pour tout $n = 2^{k+1}$ on a l'estimation

$$\| \mathbf{x}^n - \mathbf{X} \|_3 \leqslant \frac{4}{(\lambda_0^{n-1} - \lambda_0^{1-n})(\lambda_0 - \lambda_0^{-1})} \| \mathbf{x}^0 - \mathbf{X} \|_3,$$

i.e.

$$\| \mathbf{x}^n - \mathbf{X} \|_3 = \| \mathbf{x}^0 - \mathbf{X} \|_3 \, O(\lambda_0^{-n}).$$

Considérons un problème type de physique mathématique qui se ramène à résoudre un système d'équations de rapport M/μ grand. On demande de résoudre dans le carré $0 \leqslant x_1, \ x_2 \leqslant 1$ l'équation de Poisson $-\Delta u = f$ avec des conditions aux limites nulles. Donnons-nous un réseau de pas $h = 1/l$ et écrivons le système d'équations approximant le problème différentiel:

$$-\frac{u_{j_1+1, j_2} - 2u_{j_1, j_2} + u_{j_1-1, j_2}}{h^2} - \frac{u_{j_1, j_2+1} - 2u_{j_1, j_2} + u_{j_1, j_2-1}}{h^2} = f_{j_1, j_2} \qquad (20)$$

avec $0 < j_1, \ j_2 < l$;

$$u_{j_1, j_2} = 0 \quad \text{si} \quad j_1(l-j_1)\, j_2(l-j_2) = 0.$$

La matrice de ce système est définie positive et pour elle

$$\mu = \min \lambda_i = 8l^2 \sin^2\left(\frac{\pi h}{2}\right) \sim 2\pi^2,$$

$$M = \max \lambda_i = 8l^2 \cos^2\left(\frac{\pi h}{2}\right) \sim 8l^2,$$

c'est-à-dire $\sqrt{\frac{M}{\mu}} \sim \frac{2l}{\pi}$. En utilisant le pas $l = 30$, on constate un gain assez sensible sur le nombre d'itérations.

Au début du paragraphe, nous avons dit que la symétrisation d'un système risque d'en détériorer les propriétés. En effet, en appliquant ce processus à une matrice A symétrique, on passe du système $A\mathbf{x} = \mathbf{b}$ à $A^2\mathbf{x} = A\mathbf{b}$ pour lequel le rapport de la plus grande valeur propre à la plus petite n'est plus M/μ mais $(M/\mu)^2$ et la vitesse de convergence diminue en conséquence.

NOTE. Vu (19), on a pour $n \to \infty$

$$\omega_n = \frac{T_n\left(\frac{M+\mu}{M-\mu}\right)}{T_{n+1}\left(\frac{M+\mu}{M-\mu}\right)} \to \frac{1}{\lambda_0}.$$

Donc, pour n grands, la formule (16) est voisine de

$$z^{n+1} = z^n + \lambda_0^{-2}(z^n - z^{n-1}) - \frac{2}{M+\mu}(1+\lambda_0^{-2})(Az^n - b). \tag{21}$$

Si l'on itère par cette formule avec n quelconque, sous la condition $z^0 = x^0$, $z^1 = y^1$, ce processus exige approximativement le même nombre d'itérations que (16).

§ 8. Méthode de Seidel

Soit un système d'équations $A\mathbf{x} = \mathbf{b}$ dont tous les éléments diagonaux sont non nuls. En le résolvant par la méthode de Seidel on calcule successivement les composantes de la solution, la k-ième composante étant trouvée par la k-ième équation. Ainsi, si $\mathbf{x}^n = (x_1^n, \ldots, x_m^n)^T$, l'approximation suivante est donnée par le système

$$\begin{aligned} a_{11}x_1^{n+1} + a_{12}x_2^n \quad + \ldots + a_{1m}x_m^n &= b_1, \\ a_{21}x_1^{n+1} + a_{22}x_2^{n+1} + a_{23}x_3^n \quad + \ldots + a_{2m}x_m^n &= b_2, \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots & \\ a_{m1}x_1^{n+1} + a_{m2}x_2^{n+1} + \ldots + a_{mm}x_m^{n+1} &= b_m. \end{aligned} \tag{1}$$

Le système (1) se récrit :

$$B\mathbf{x}^{n+1} + C\mathbf{x}^n = \mathbf{b} \tag{2}$$

avec

$$B = \begin{pmatrix} a_{11} & 0 & 0 & \ldots & 0 \\ a_{21} & a_{22} & 0 & \ldots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ a_{m1} & a_{m2} & a_{m3} & \ldots & a_{mm} \end{pmatrix},$$

$$C = \begin{pmatrix} 0 & a_{12} & a_{13} & \ldots & a_{1m} \\ 0 & 0 & a_{23} & \ldots & a_{2m} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & \ldots & 0 \end{pmatrix}.$$

D'où

$$\mathbf{x}^{n+1} = -B^{-1}C\mathbf{x}^n + B^{-1}\mathbf{b}. \tag{3}$$

La méthode de Seidel équivaut donc à une méthode itérative simple, et pour qu'elle converge, quelle que soit l'approximation de départ,

il faut et il suffit que toutes les valeurs propres de la matrice $-B^{-1}C$ aient un module inférieur à 1. L'égalité

$$\det(-B^{-1}C - \lambda E) = \det(-B^{-1}) \det(C + B\lambda)$$

entraîne que les valeurs propres de $(-B^{-1}C)$ sont racines de l'équation $\det(C + B\lambda) = 0$.

Donc, la condition nécessaire et suffisante de convergence de la méthode de Seidel est que toutes les racines de l'équation

$$\det\begin{pmatrix} a_{11}\lambda & a_{12} & a_{13} & \dots & a_{1m} \\ a_{21}\lambda & a_{22}\lambda & a_{23} & \dots & a_{2m} \\ \dots & \dots & \dots & \dots & \dots \\ a_{m1}\lambda & a_{m2}\lambda & a_{m3}\lambda & \dots & a_{mm}\lambda \end{pmatrix} = 0$$

aient un module plus petit que l'unité.

Nous essayerons d'obtenir des critères de suffisance plus commodes pour les applications.

THÉORÈME. *Soit*

$$\sum_{j \neq i} |a_{ij}| \leqslant q\,|a_{ii}|, \quad q < 1,$$

quel que soit i. Alors

$$\|\mathbf{x}^{n+1} - \mathbf{X}\|_1 \leqslant q\,\|\mathbf{x}^n - \mathbf{X}\|_1.$$

DÉMONSTRATION. Retranchant de (2) l'égalité $B\mathbf{X} + C\mathbf{X} = \mathbf{b}$ vérifiée par la solution exacte, on obtient l'équation d'erreur

$$B\mathbf{r}^{n+1} + C\mathbf{r}^n = 0. \tag{4}$$

Soit $\|\mathbf{r}^{n+1}\|_1 = \|r_l^{n+1}\|$; écrivons la l-ième équation du système (4)

$$\sum_{j=1}^{l-1} a_{lj} r_j^{n+1} + a_{ll} r_l^{n+1} + \sum_{j=l+1}^{m} a_{lj} r_j^n = 0,$$

et résolvons-la par rapport à r_l^{n+1} :

$$r_l^{n+1} = -\sum_{j=1}^{l-1} \frac{a_{lj}}{a_{ll}} r_j^{n+1} - \sum_{j=l+1}^{m} \frac{a_{lj}}{a_{ll}} r_j^n.$$

Il en résulte

$$|r_l^{n+1}| \leqslant \alpha\,\|\mathbf{r}^{n+1}\|_1 + \beta\,\|\mathbf{r}^n\|_1 \tag{5}$$

avec

$$\alpha = \sum_{j=1}^{l-1} \left|\frac{a_{lj}}{a_{ll}}\right|, \qquad \beta = \sum_{j=l+1}^{m} \left|\frac{a_{lj}}{a_{ll}}\right|;$$

une autre écriture de cette relation est

$$\| r^{n+1} \|_1 \leqslant \alpha \| r^{n+1} \|_1 + \beta \| r^n \|_1$$

d'où

$$\| r^{n+1} \|_1 \leqslant \frac{\beta}{1-\alpha} \| r^n \|_1.$$

Par hypothèse $\alpha + \beta \leqslant q < 1$; c'est pourquoi

$$\frac{\beta}{1-\alpha} \leqslant \frac{q-\alpha}{1-\alpha} = q - \frac{\alpha(1-q)}{1-\alpha} \leqslant q$$

et

$$\| r^{n+1} \|_1 \leqslant \frac{\beta}{1-\alpha} \| r^n \|_1 \leqslant q \| r^n \|_1.$$

Signalons certaines circonstances qui pourront nous être utiles lors de l'étude de la convergence de la méthode de Seidel. La multiplication des relations (1) par certains facteurs $\alpha^i \neq 0$ laisse invariants les vecteurs x^n définis par ces relations. La convergence de la méthode de Seidel est conservée donc si l'on multiplie les lignes de la matrice A par des nombres. Si l'on multiplie la j-ième colonne de A par $\beta_j \neq 0$ et divise x_j^0 par β_j, toutes les quantités x_j^n sont divisées elles aussi par β_j. Ainsi, la convergence de notre méthode se conserve par multiplication des colonnes de la matrice A par des nombres quelconques.

La résolution des systèmes d'équations linéaires est un modèle sur lequel on étudie les problèmes relativement plus compliqués que sont la résolution des systèmes d'équations non linéaires et la

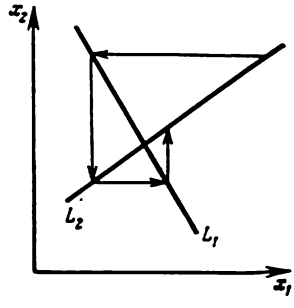


Fig. 6.8.1

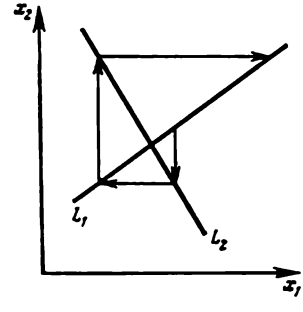


Fig. 6.8.2

minimisation des fonctions de plusieurs variables. On ne saurait transporter une méthode à des problèmes plus ardus sans en comprendre les propriétés qualitatives, les plus grossières qui garantissent la convergence. A cette fin, la méthode doit avant tout être interprétée géométriquement. Désignons par L_i le plan $\sum_{j=1}^{m} a_{ij}x_j - b_i = 0$. La valeur itérée $(x_1^{n+1}, \ldots, x_i^{n+1}, x_{i+1}^n, \ldots, x_m^n)$ se déduit de

$(x_1^{n+1}, \ldots, x_{i-1}^{n+1}, x_i^n, \ldots, x_m^n)$ par un déplacement de l'approximation parallèlement à l'axe des x_i jusqu'à l'intersection avec le plan L_i; géométriquement, la méthode de Seidel consiste donc en un déplacement cyclique parallèle aux axes des x_i jusqu'à la rencontre avec L_i. Les fig. 6.8.1-6.8.3 illustrent, pour $m = 2$, les cas

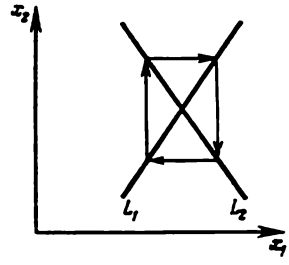


Fig. 6.8.3

où la méthode de Seidel converge, diverge ou cycle. En comparant les deux premiers dessins on voit que la convergence peut changer de caractère par permutation des équations.

Le cas de A symétrique a un support géométrique particulièrement intéressant.

THÉORÈME. *Si A est une matrice réelle, symétrique et définie positive, la méthode de Seidel converge.*

DÉMONSTRATION. On a, lorsque A est symétrique,

$$F(\mathbf{y}) = (A(\mathbf{y}-\mathbf{X}),\ \mathbf{y}-\mathbf{X}) - (A\mathbf{X}, \mathbf{X}) =$$

$$= (A\mathbf{y}, \mathbf{y}) - 2(A\mathbf{X}, \mathbf{y}) = (A\mathbf{y}, \mathbf{y}) - 2(\mathbf{b}, \mathbf{y}).$$

Si $A > 0$, alors $(A(\mathbf{y} - \mathbf{X}), \mathbf{y} - \mathbf{X}) > 0$, $\mathbf{y} \neq \mathbf{X}$, et la fonction $F(\mathbf{y})$ admet un minimum et un seul pour $\mathbf{y} = \mathbf{X}$. Résoudre le système $A\mathbf{x} = \mathbf{b}$ revient donc à chercher l'unique minimum de $F(\mathbf{y})$.

Citons, parmi les techniques de minimisation des fonctions de plusieurs variables, la *méthode de descente séquentielle.*

Soit $(x_1^0, \ldots, x_m^0)$ une approximation du point de minimum de la fonction $F(x_1, \ldots, x_m)$. Considérons la fonction $F(x_1, x_2^0, \ldots, x_m^0)$ en tant que fonction de la variable x_1 et trouvons le point x_1^1 réalisant son minimum. Prenons maintenant pour point de départ l'approximation $(x_1^1, x_2^0, \ldots, x_m^0)$ obtenue en minimisant $F(x_1^1, x_2, x_3^0, \ldots, x_m^0)$ et trouvons l'approximation suivante $(x_1^1, x_2^1, x_3^0, \ldots, x_m^0)$. Le processus réitère de façon cyclique. Lors de l'amélioration des valeurs approchées de x_k, on se déplace sur une droite parallèle à l'axe des x_k jusqu'au point réalisant la valeur

$F(x) = c$ la plus petite possible pour cette droite. Il s'agit évidemment du point de contact de ladite droite et de la ligne de niveau $F(x) = c$. Aussi, dans le cas bidimensionnel, le processus se présente comme sur la fig. 6.8.4.

Appliquons la méthode de descente séquentielle pour obtenir l'extrémum de la fonction F. Désignons $F(y) + (AX, X) = (A(y - X), y - X)$ par $F_0(y)$. En minimisant en x_k on se

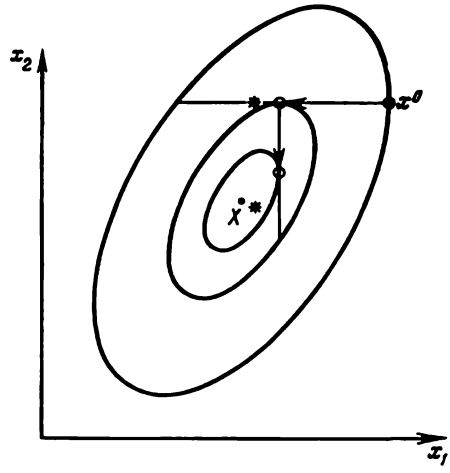


Fig. 6.8.4

déplace parallèlement à l'axe des x_k jusqu'au point en lequel $F'_{x_k} = 0$. Par conséquent, la nouvelle valeur x_k se définit à l'aide de

$$F'_{x_k} = 2\left(\sum_{j=1}^{m} a_{kj}x_j - b_k\right) = 0,$$

i.e. par la même équation que dans la méthode de Seidel. Ainsi, en minimisant la fonction F par la méthode de descente séquentielle et en résolvant le système initial par la méthode de Seidel, on obtient les mêmes approximations.

Si $x^n \neq X$, l'une des équations au moins n'est pas satisfaite, et la valeur correspondante $F'_{x_k}(x^n)$ est différente de zéro. Choisissons le plus petit de ces k. Alors, en approchant $x_1, \ldots, x_{k-1}$ nous restons en x^n et en approchant x_k on se déplace vers $F(x)$ les plus petits; en précisant les autres composantes, la valeur $F(x)$ ne croît pas. Donc.

$$F(x^{n+1}) < F(x^n), \quad F_0(x^{n+1}) < F_0(x^n),$$

ce qui implique

$$\frac{F_0(x^{n+1})}{F_0(x^n)} < 1 \tag{6}$$

pour $x^n \neq X$. Par suite de $(4)_2$ on a $r^{n+1} = -B^{-1}Cr^n$. La relation (6) s'écrit

$$\varphi(r^n) = \frac{(AB^{-1}Cr^n,\ B^{-1}Cr^n)}{(Ar^n,\ r^n)} < 1 \tag{7}$$

avec $r^n \neq 0$. Sur la sphère $\| r^n \| = 1$, $\varphi(r^n)$ est continue et atteint donc sa plus grande valeur φ_0. Puisque $A > 0$, on a toujours $\varphi(r^n) > 0$, donc $\varphi_0 > 0$; posons $\sqrt{\varphi_0} = \lambda$. Vu (7), on a $\lambda^2 < 1$. Evidemment $\varphi(cr^n) = \varphi(r^n)$ quel que soit $c \neq 0$; aussi $\varphi(r^n) = \varphi(r^n/\| r^n \|) \leqslant \lambda^2$ avec r^n quelconque. D'où l'inégalité

$$\frac{F_0(x^{n+1})}{F_0(x^n)} \leqslant \lambda^2, \tag{8}$$

et

$$F_0(x^n) \leqslant \lambda^{2n} F_0(x^0).$$

(3.8) et (3.9) entraînent

$$(\min \lambda_A^i) \| y - X \|_3^2 \leqslant F_0(y) \leqslant (\max \lambda_A^i) \| y - X \|_3^2,$$

ce qui permet d'évaluer la vitesse de convergence:

$$\| x^n - X \|_3 \leqslant$$

$$\leqslant \sqrt{\frac{F_0(x^n)}{\min \lambda_A^i}} \leqslant \lambda^n \sqrt{\frac{F_0(x^0)}{\min \lambda_A^i}} \leqslant \lambda^n \sqrt{\frac{\max \lambda_A^i}{\min \lambda_A^i}} \| x^0 - X \|_3. \tag{9}$$

Le théorème est démontré.

La fig. 6.8.5 montre que la méthode de Seidel converge plus vite si la direction des axes des ellipsoïdes est proche de celle des axes de coordonnées, c'est-à-dire si la matrice A est proche de la forme diagonale.

Dans le cas 6.8.4 les valeurs approchées successives se déplacent de manière monotone à gauche et vers le bas. Un déplacement monotone de certaines composantes dans une même direction est propre à plusieurs classes matricielles. Ce déplacement caractérise assez souvent les composantes dont la vitesse de convergence est la plus mauvaise. Pour l'accélérer on recourt alors à la *méthode de relaxation*. Après avoir calculé chaque coordonnée par la méthode de Seidel on effectue, dans la même direction, un déplacement supplémentaire égal à la p-ième partie du déplacement précédent (p non obligatoirement positif). Les approximations de la méthode de relaxation se déterminent donc par la relation

$$B\left(\frac{x^{n+1} + px^n}{1+p}\right) + Cx^n = b. \tag{10}$$

La pratique des calculs montre que pour $A > 0$ on prendra le facteur de relaxation p dans les limites $0 < p < 1$; on a alors affaire à la *surrelaxation*. Les petits, cercles blancs de la fig. 6.8.4 représentent les approximations de la méthode de Seidel et les astérisques celles de la surrelaxation pour $p = 1/4$. En résolvant le système (7.20) par la technique de Seidel une réduction de l'erreur de ε^{-1} fois exige $\sim ch^{-2} \ln(\varepsilon^{-1})$ itérations, tandis que la méthode de surrelaxation, pour $p = 1 - \omega h$, $\omega > 0$, n'en demande que $\sim c(\omega) h^{-1} \ln(\varepsilon^{-1})$. Le choix du paramètre ω est en l'occurrence très délicat. En particulier, dans nombre de cas la valeur optimale de p se définit expérimentalement.

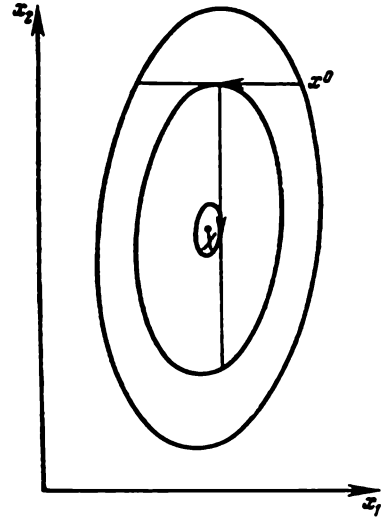


Fig. 6.8.5

On choisit parfois le facteur de relaxation p en fonction de n et i.

Revenons de nouveau à la fig. 6.8.4 pour $A > 0$. Après avoir approché la composante x_1 par la méthode de surrelaxation avec $-1 < p < 1$, nous tombons en un point situé à l'intérieur ou sur la frontière de l'ellipsoïde $F(\mathbf{x}) = F(\mathbf{x}^0)$. En nous rappelant nos raisonnements relatifs à la méthode de Seidel, nous tirons la conclusion que pour p dans les limites indiquées on a toujours $F_0(\mathbf{x}^{n+1}) < F_0(\mathbf{x}^n)$ si $\mathbf{x}^n \neq \mathbf{X}$.

PROBLÈME. Démontrer que dans la condition $|p_n| \leqslant q < 1$ la méthode de relaxation converge en progression géométrique.

§ 9. Méthode de la plus grande pente

Un procédé répandu de minimisation des fonctions de plusieurs variables est la *méthode du gradient*. Chaque approximation s'obtient de la précédente par un déplacement dans la direction du gradient de $F(\mathbf{x})$ et s'écrit donc

$$\mathbf{x}^{n+1} = \mathbf{x}^n - \delta_n \operatorname{grad} F(\mathbf{x}^n). \tag{1}$$

Cette description ne définit pas l'algorithme de façon unique puisqu'elle ne dit rien sur le choix du paramètre δ_n. Celui-ci se détermine par exemple par la condition de minimum de

$$F(\mathbf{x}^n - \delta_n \operatorname{grad} F(\mathbf{x}^n)). \tag{2}$$

Il s'agit dans ce cas de la *méthode de la plus grande pente.*

Pour la fonction $F(\mathbf{x}) = (A\mathbf{x}, \mathbf{x}) - 2(\mathbf{b}, \mathbf{x})$ associée au système d'équations linéaires de matrice $A > 0$ le problème de recherche du minimum est un problème qui se résout explicitement.

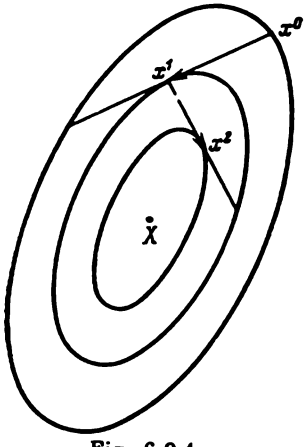


Fig. 6.9.1

Dans ce cas concret

$$\operatorname{grad} F = 2(A\mathbf{x} - \mathbf{b})$$

et

$$\mathbf{x}^{n+1} = \mathbf{x}^n - \delta_n 2(A\mathbf{x}^n - \mathbf{b}).$$

Désignons $2\delta_n$ par Δ_n, i.e. posons

$$\mathbf{x}^{n+1} = \mathbf{x}^n - \Delta_n(A\mathbf{x}^n - \mathbf{b}). \tag{3}$$

Soit $\varphi(\Delta_n) = F(\mathbf{x}^{n+1})$. Rappelons-nous que $A = A^*$ et calculons $\varphi'(\Delta_n)$:

$$\varphi'(\Delta_n) = \left(A\frac{d\mathbf{x}^{n+1}}{d\Delta_n},\ \mathbf{x}^{n+1}\right) + \left(A\mathbf{x}^{n+1},\ \frac{d\mathbf{x}^{n+1}}{d\Delta_n}\right) - 2\left(\frac{d\mathbf{x}^{n+1}}{d\Delta_n},\ \mathbf{b}\right) =$$

$$= 2\left(A\mathbf{x}^{n+1} - \mathbf{b},\ \frac{d\mathbf{x}^{n+1}}{d\Delta_n}\right) = -2(A\mathbf{x}^{n+1} - \mathbf{b},\ A\mathbf{x}^n - \mathbf{b}).$$

Compte tenu de (3), l'équation $\varphi'(\Delta_n) = 0$ s'écrit

$$(A\mathbf{x}^n - \mathbf{b} - \Delta_n A(A\mathbf{x}^n - \mathbf{b}),\ A\mathbf{x}^n - \mathbf{b}) = 0,$$

d'où

$$\Delta_n = \frac{(A\mathbf{x}^n - \mathbf{b},\ A\mathbf{x}^n - \mathbf{b})}{(A(A\mathbf{x}^n - \mathbf{b}),\ A\mathbf{x}^n - \mathbf{b})}. \tag{4}$$

La fig. 6.9.1 représente les approximations de la méthode de la plus grande pente et les courbes de niveau de la fonction F_0. On dit que le processus itératif (3), (4) *utilise les plus grandes pentes pour résoudre le système* considéré *d'équations linéaires.*

Soit $S_A \subset [\mu, M]$.

THÉORÈME. *Les approximations de la méthode de la plus grande pente vérifient*

$$F_0(\mathbf{x}^n) \leqslant \left(\frac{M-\mu}{M+\mu}\right)^{2n} F_0(\mathbf{x}^0). \tag{5}$$

DÉMONSTRATION. Pour $\mathbf{y}^n = \mathbf{x}^n$, effectuons une itération du processus optimal à pas séparés

$$\mathbf{y}^{n+1} = \mathbf{y}^n - \frac{2}{M+\mu}(A\mathbf{y}^n - \mathbf{b}). \tag{6}$$

Les erreurs d'itération $\mathbf{r}^n = \mathbf{y}^n - \mathbf{X}$ sont reliées par

$$\mathbf{r}^{n+1} = \left(E - \frac{2}{M+\mu} A\right)\mathbf{r}^n.$$

Soit $\mathbf{e}_1, \ldots$ un système orthonormé de vecteurs propres de la matrice A: $A\mathbf{e}_i = \lambda_i \mathbf{e}_i$.

Puisque $\mu \leqslant \lambda_i \leqslant M$, on a, quel que soit i,

$$-\frac{M-\mu}{M+\mu} \leqslant 1 - \frac{2}{M+\mu}\lambda_i \leqslant \frac{M-\mu}{M+\mu}$$

et donc

$$\left|1 - \frac{2}{M+\mu}\lambda_i\right| \leqslant \frac{M-\mu}{M+\mu}. \tag{7}$$

Soit $\mathbf{r}^n = \sum c_i \mathbf{e}_i$. On a les relations

$$(A\mathbf{r}^n, \mathbf{r}^n) = \left(\sum c_i \lambda_i \mathbf{e}_i, \sum c_i \mathbf{e}_i\right) = \sum \lambda_i c_i^2,$$

$$\mathbf{r}^{n+1} = \sum b_i \mathbf{e}_i, \quad \text{où} \quad b_i = \left(1 - \frac{2\lambda_i}{M+\mu}\right) c_i,$$

$$(A\mathbf{r}^{n+1}, \mathbf{r}^{n+1}) = \sum \lambda_i b_i^2 = \sum \lambda_i \left(1 - \frac{2\lambda_i}{M+\mu}\right)^2 c_i^2.$$

Compte tenu de (7) il vient

$$(A\mathbf{r}^{n+1}, \mathbf{r}^{n+1}) \leqslant \left(\frac{M-\mu}{M+\mu}\right)^2 (A\mathbf{r}^n, \mathbf{r}^n).$$

Vu que $F_0(\mathbf{y}^n) = (A\mathbf{r}^n, \mathbf{r}^n)$, cela signifie que

$$F_0(\mathbf{y}^{n+1}) \leqslant \left(\frac{M-\mu}{M+\mu}\right)^2 F_0(\mathbf{y}^n) = \left(\frac{M-\mu}{M+\mu}\right)^2 F_0(\mathbf{x}^n).$$

On écrit l'approximation $\mathbf{y}^{n+1}$ sous la forme (1):

$$\mathbf{y}^{n+1} = \mathbf{x}^n - \alpha \operatorname{grad} F(\mathbf{x}^n), \quad \alpha = \frac{1}{M+\mu}.$$

Etant donné que de toutes les approximations de la forme (1) c'est $\mathbf{x}^{n+1}$ qui réalise le minimum de la fonctionnelle $F(\mathbf{x})$, on a

$$F(\mathbf{x}^{n+1}) \leqslant F(\mathbf{y}^{n+1}).$$

Il en résulte l'estimation

$$F_0(\mathbf{x}^{n+1}) \leqslant F_0(\mathbf{y}^{n+1}) \leqslant \left(\frac{M-\mu}{M+\mu}\right)^2 F_0(\mathbf{x}^n).$$

et, partant, le théorème qu'on avait à démontrer. De même que (8.9) on obtient l'inégalité

$$\|\mathbf{x}^n - \mathbf{X}\|_3 \leqslant \left(\frac{M-\mu}{M+\mu}\right)^n \sqrt{\frac{M}{\mu}}\, \|\mathbf{x}^0 - \mathbf{X}\|_3.$$

Bien qu'à chaque pas la quantité $F_0(x)$ ne diminue pas moins dans la méthode de la plus grande pente que dans le processus itératif (6), les estimations de la vitesse de convergence sont quasi identiques. Ces procédés et évaluations n'en présentent pas moins une différence de principe. En effet, le processus itératif (6) a besoin pour se dérouler d'une information sur les bornes μ, M du spectre, tandis que la méthode (3)-(4) s'en passe parfaitement.

Le procédé utilisant la plus grande pente présente le handicap suivant par rapport au processus simple (6). C'est que chacune de ses approximations s'obtient par deux opérations pénibles (au lieu d'une) de multiplication de la matrice par le vecteur.

Pour éviter cet inconvénient on note $\mathbf{w}^n = A\mathbf{x}^n - \mathbf{b}$ et on écrit (3) sous la forme

$$\mathbf{x}^{n+1} = \mathbf{x}^n - \Delta_n \mathbf{w}^n. \tag{8}$$

Multipliant à gauche par A et retranchant $\mathbf{b}$, il vient

$$\mathbf{w}^{n+1} = \mathbf{w}^n - \Delta_n A\mathbf{w}^n. \tag{9}$$

La formule (4) donnant Δ_n peut se récrire

$$\Delta_n = \frac{(\mathbf{w}^n, \mathbf{w}^n)}{(A\mathbf{w}^n, \mathbf{w}^n)}. \tag{10}$$

En opérant par itérations on met en réserve les vecteurs $\mathbf{x}^n$, $\mathbf{w}^n$ et on calcule successivement à chaque pas $A\mathbf{w}^n$, Δ_n, $\mathbf{x}^{n+1}$, $\mathbf{w}^{n+1}$. Dans la méthode initiale (3)-(4) de la plus grande pente l'erreur par pas équivaut à une perturbation de l'approximation de départ et, le processus étant convergent, son influence doit avoir tendance à s'amortir. Le processus (8)-(10) donne lieu à une relation plus compliquée entre les erreurs sur les approximations successives, qui entraîne souvent un accroissement notable de l'erreur de calcul totale avec n. Nous conseillons donc d'interrompre de temps en temps les calculs par les formules (8)-(10) et de faire une itération selon (3)-(4). L'erreur d'arrondi commise lors des pas précédents devient alors, pour les opérations suivantes, l'erreur sur la condition initiale et a tendance à diminuer.

On choisira tel ou tel processus itératif en se guidant sur l'information que l'on possède sur les bornes du spectre, le volume et la structure de la mémoire de la machine à calculer électronique. Ainsi, en approximant sur un réseau des équations aux dérivées partielles, on procède parfois comme suit. On effectue des calculs supplémentaires visant à déterminer le plus exactement possible les valeurs μ et M, puis on applique le processus itératif linéaire optimal.

§ 10. Méthode du gradient conjugué

Cette méthode s'applique aux systèmes d'équations

$$A\mathbf{x} = \mathbf{b}$$

de matrice A définie positive.

Soit $\mathbf{x}^0$ l'approximation initiale de la solution. Formons le vecteur

$$\mathbf{x}^n = \mathbf{x}^0 + \sum_{i=0}^{n-1} \alpha_i A^i \mathbf{r}^0, \quad \mathbf{r}^0 = A\mathbf{x}^0 - \mathbf{b} \tag{1}$$

et proposons-nous de minimiser

$$F(\mathbf{x}^n) = (A\mathbf{x}^n, \mathbf{x}^n) - 2(\mathbf{b}, \mathbf{x}^n)$$

moyennant le choix des paramètres α_i. On a

$$\frac{dF(\mathbf{x}^n)}{d\alpha_j} = 2\left(A\mathbf{x}^n - \mathbf{b}, \frac{d\mathbf{x}^n}{d\alpha_j}\right),$$

$\frac{d\mathbf{x}^n}{d\alpha_j} = A^j\mathbf{r}_0$ et

$$A\mathbf{x}^n - \mathbf{b} = \mathbf{r}^0 + \sum_{i=0}^{n-1} \alpha_i A^{i+1}\mathbf{r}^0, \tag{2}$$

les conditions $F'_{\alpha_j} = 0$ forment le système

$$(A\mathbf{x}^n - \mathbf{b}, A^j\mathbf{r}^0) = 0, \quad j = 0, \ldots, n-1. \tag{3}$$

En y portant l'expression de $A\mathbf{x}^n - \mathbf{b}$, on aboutit à un système d'équations linéaires

$$(\mathbf{r}^0, A^j\mathbf{r}^0) + \sum_{i=0}^{n-1} \alpha_i (A^{i+1}\mathbf{r}^0, A^j\mathbf{r}^0) = 0, \quad j = 0, \ldots, n-1, \tag{4}$$

en $\alpha_0, \ldots, \alpha_{n-1}$.

Si $\mathbf{x}^0$ est donnée, on peut choisir un certain n, résoudre le système (4) en α_i et obtenir l'approximation $\mathbf{x}^n$. L'algorithme décrit connaît plusieurs utilisations.

Ainsi, on peut opérer de proche en proche en adoptant en qualité d'approximation initiale à chaque pas l'approximation obtenue au pas précédent.

On peut également construire une suite de vecteurs

$$\mathbf{x}^n = \mathbf{x}^0 + \sum_{i=0}^{n-1} \alpha_i^n A^i (A\mathbf{x}^0 - \mathbf{b}),$$

qui minimisent $F(\mathbf{x})$ pour chaque n.

Proposons-nous d'établir une relation de récurrence qui lie ces vecteurs à la manière des relations du processus itératif linéaire optimal.

Soit $\mathbf{e}_1, \ldots, \mathbf{e}_m$ un système orthonormé complet de vecteurs propres de la matrice A. Mettons le vecteur $\mathbf{r}^0 = A\mathbf{x}^0 - \mathbf{b}$ sous la forme $\mathbf{r}^0 = \sum_{j=1}^{m} c_j\mathbf{e}_j$. Une combinaison linéaire de vecteurs propres correspondant à une même valeur propre est encore vecteur propre associé à cette valeur propre. Aussi, en groupant les termes correspondant aux valeurs propres identiques et en chassant les termes nuls, on obtient $\mathbf{r}^0 = \sum_{j=1}^{q} \mathbf{g}_j$; ici $\mathbf{g}_j \neq 0$,

$$A\mathbf{g}_j = \lambda_j\mathbf{g}_j,$$

$\lambda_1, \ldots, \lambda_q$ étant des valeurs propres distinctes de la matrice A. Puisque les vecteurs propres associés à des valeurs propres différentes sont orthogonaux, on a $(\mathbf{g}_i, \mathbf{g}_j) = 0$ lorsque $i \neq j$.

Montrons l'indépendance linéaire des vecteurs $\mathbf{r}^0, \ldots, A^{q-1}\mathbf{r}^0$. Supposons le contraire; alors $\sum_{k=0}^{q-1} \gamma_k A^k \mathbf{r}^0 = 0$, ou encore

$$0 = \sum_{k=0}^{q-1} \gamma_k \Big(\sum_{j=1}^{q} (\lambda_j)^k \mathbf{g}_j \Big) = \sum_{j=1}^{q} P(\lambda_j)\, \mathbf{g}_j$$

avec $P(\lambda) = \sum_{k=0}^{q-1} \gamma_k \lambda^k$. L'indépendance linéaire des $\mathbf{g}_j$ entraîne

$$P(\lambda_j) = 0, \quad j = 1, \ldots, q.$$

Contradiction: un polynôme de degré plus petit que q a q zéros distincts.

Il en résulte que le vecteur

$$A\mathbf{x}^n - \mathbf{b} = \mathbf{r}^0 + \sum_{i=1}^{n-1} \alpha_i^n A^i \mathbf{r}^0$$

est différent de zéro pour $n < q$.

Montrons que $\mathbf{x}^q = \mathbf{X}$. En effet, définissons le polynôme

$$Q_q(\lambda) = \prod_{j=1}^{q} \Big(1 - \frac{\lambda}{\lambda_j}\Big) = 1 + \sum_{k=1}^{q} \beta_k \lambda^k.$$

Si

$$\bar{\mathbf{x}}^q = \mathbf{x}^0 + \sum_{k=1}^{q} \beta_k A^{k-1} (A\mathbf{x}^0 - \mathbf{b}),$$

alors, d'après (2) et la définition ci-dessus, nous avons

$$A\bar{\mathbf{x}}^q - \mathbf{b} = \sum_{j=1}^{q} Q_q(\lambda_j)\, \mathbf{g}_j = 0;$$

donc $\bar{\mathbf{x}}^q$ réalise un minimum de la fonctionnelle $F(\mathbf{x})$ et, par conséquent, $\bar{\mathbf{x}}^q = \mathbf{x}^q = \mathbf{X}$.

LEMME. *α_{n-1}^n est différent de zéro pour $n \leqslant q$.*

DÉMONSTRATION. Si $\alpha_{q-1}^q = 0$, il vient

$$0 = A\mathbf{x}^q - \mathbf{b} = \mathbf{r}^0 + \sum_{i=0}^{q-2} \alpha_i^q A^{i+1} \mathbf{r}^0,$$

ce qui contredit l'hypothèse des $\mathbf{r}^0, \ldots, A^{q-1}\mathbf{r}^0$ linéairement indépendants. Soit $\alpha_{j-1}^j = 0$ et $j < q$; le vecteur $A\mathbf{x}^j - \mathbf{b}$ s'écrit, en vertu

de (2), sous forme d'une combinaison linéaire de $\mathbf{r}^0, \ldots, A^{j-1}\mathbf{r}^0$; de plus, vu (3), il est orthogonal à tous ces derniers. D'où $A\mathbf{x}^j - \mathbf{b} = 0$ et, donc, $j \geqslant q$. Le lemme est démontré.

Soit l'ensemble des égalités

$$\begin{aligned}
\mathbf{x}^1 &= \mathbf{x}^0 + \alpha_0^1 \mathbf{r}^0, \\
\mathbf{x}^2 &= \mathbf{x}^0 + \alpha_0^2 \mathbf{r}^0 + \alpha_1^2 A\mathbf{r}^0, \\
&\ldots\ldots\ldots\ldots\ldots \\
\mathbf{x}^q &= \mathbf{x}^0 + \ldots + \alpha_{q-1}^q A^{q-1}\mathbf{r}^0.
\end{aligned}$$

Etant donné la condition $\alpha_{j-1}^j \neq 0$, ce système de matrice triangulaire se résout par rapport aux inconnues $A^j\mathbf{r}^0$, $j = 0, \ldots, q-1$; il en résulte les relations

$$A^{j-1}\mathbf{r}^0 = \sum_{i=1}^{j} b_i^j (\mathbf{x}^i - \mathbf{x}^0), \quad j = 1, \ldots, q.$$

Portons les expressions de $A^{j-1}\mathbf{r}^0$ dans le deuxième membre de (2); il vient

$$A\mathbf{x}^n - \mathbf{b} = \mathbf{r}^0 + \sum_{i=0}^{n-1} \alpha_i^n A^{i+1}\mathbf{r}^0 = \sum_{i=1}^{n+1} \gamma_i^n (\mathbf{x}^i - \mathbf{x}^0) \quad \text{pour} \quad n < q.$$

Faisons agir l'opérateur A sur les deux membres et obtenons

$$A(A\mathbf{x}^n - \mathbf{b}) = \sum_{i=1}^{n+1} \gamma_i^n ((A\mathbf{x}^i - \mathbf{b}) - (A\mathbf{x}^0 - \mathbf{b})) = \sum_{i=0}^{n+1} \sigma_i^n (A\mathbf{x}^i - \mathbf{b}). \qquad (5)$$

L'égalité $\sum_{i=0}^{n+1} \sigma_i^n = 0$ est évidente. Les relations (3) se récrivent

$$(A(A\mathbf{x}^n - \mathbf{b}), A^j\mathbf{r}^0) = 0, \quad j = 0, \ldots, n-2;$$

compte tenu de (2), il en résulte

$$(A(A\mathbf{x}^n - \mathbf{b}), A\mathbf{x}^k - \mathbf{b}) = 0 \quad \text{pour} \quad k \leqslant n-2.$$

De même

$$(A\mathbf{x}^n - \mathbf{b}, A\mathbf{x}^k - \mathbf{b}) = 0 \quad \text{pour} \quad k \leqslant n-1.$$

En multipliant scalairement (5) par les vecteurs $(A\mathbf{x}^k - \mathbf{b})$, $k = 0, \ldots, n-2$, on trouve

$$0 = \sigma_k^n (A\mathbf{x}^k - \mathbf{b}, A\mathbf{x}^k - \mathbf{b}),$$

donc tous les σ_k^n sont nuls lorsque $k < n-1$. L'égalité (5) prend alors la forme suivante:

$$A(A\mathbf{x}^n - \mathbf{b}) = \sum_{i=n-1}^{n+1} \sigma_i^n (A\mathbf{x}^i - \mathbf{b});$$

l'application de l'opérateur A^{-1} donne

$$A\mathbf{x}^n - \mathbf{b} = \sum_{i=n-1}^{n+1} \sigma_i^n (\mathbf{x}^i - \mathbf{X}).$$

Comme $\sum \sigma_i^n = 0$, cette dernière égalité peut s'écrire

$$A\mathbf{x}^n - \mathbf{b} = \sigma_{n-1}^n \mathbf{x}^{n-1} + \sigma_n^n \mathbf{x}^n + \sigma_{n+1}^n \mathbf{x}^{n+1}. \tag{6}$$

On remarquera que les constructions que nous venons d'effectuer rappellent beaucoup celles du § 15, ch. II.

Pour les valeurs considérées $n < q$, le coefficient σ_{n+1}^n est toujours non nul. Démontrons-le par l'absurde. La substitution de

$$A\mathbf{x}^n - \mathbf{b} = \mathbf{r}^0 + \sum_{i=0}^{n-1} \alpha_i^n A^{i+1} \mathbf{r}^0,$$

$$\mathbf{x}^k - \mathbf{x}^0 = \sum_{i=0}^{k-1} \alpha_i^k A^i \mathbf{r}^0, \quad k = n-1,\ n,$$

dans

$$A\mathbf{x}^n - \mathbf{b} - \sum_{i=n-1}^{n+1} \sigma_i^n (\mathbf{x}^i - \mathbf{x}^0) = 0$$

donne alors

$$\alpha_{n-1}^n A^n \mathbf{r}^0 + \ldots = 0,$$

ce qui est contraire à l'hypothèse de l'indépendance linéaire des vecteurs $\mathbf{r}^0, \ldots, A^n\mathbf{r}^0$. En effectuant le produit scalaire de (5) par $A\mathbf{x}^k - \mathbf{b}$, $k = n-1,\ n$, on trouve, pour les coefficients σ_{n-1}^n et σ_n^n :

$$\sigma_k^n = \frac{(A(A\mathbf{x}^n - \mathbf{b}),\ A\mathbf{x}^k - \mathbf{b})}{(A\mathbf{x}^k - \mathbf{b},\ A\mathbf{x}^k - \mathbf{b})},$$

et puisque $A\mathbf{x}^k - \mathbf{b} \neq 0$ et $A > 0$, il vient $\sigma_{n-1}^n,\ \sigma_n^n \neq \infty$. Comme $\sigma_{n+1}^n = -\sigma_n^n - \sigma_{n-1}^n$ on peut écrire (6) sous la forme

$$A\mathbf{x}^n - \mathbf{b} = \sigma_n^n (\mathbf{x}^n - \mathbf{x}^{n+1}) + \sigma_{n-1}^n (\mathbf{x}^{n-1} - \mathbf{x}^{n+1}). \tag{7}$$

Déterminant σ_n^n et σ_{n-1}^n pour chaque n, on trouve de proche en proche à l'aide de (7) les vecteurs $\mathbf{x}^2, \ldots$.

Le vecteur $\mathbf{x}^1$ se définit immédiatement en résolvant le système (4) : il coïncide avec la première approximation dans la méthode de la plus grande pente. Afin de trouver $\mathbf{x}^{n+1}$ à partir de (7) il faut deux opérations de multiplication de la matrice par le vecteur : $A\mathbf{x}^n$ et $A(A\mathbf{x}^n)$, soit $\sim 2m^2$ opérations arithmétiques chacune, et le calcul de plusieurs produits scalaires. Ce processus itératif est donc environ deux fois plus laborieux que le processus linéaire optimal par itérations. Le mode opératoire se simplifie de la manière suivante. L'égalité (7) entraîne

$$A(A\mathbf{x}^n - \mathbf{b}) = \sigma_n^n ((A\mathbf{x}^n - \mathbf{b}) - (A\mathbf{x}^{n+1} - \mathbf{b})) +$$

$$+ \sigma_{n-1}^n ((A\mathbf{x}^{n-1} - \mathbf{b}) - (A\mathbf{x}^{n+1} - \mathbf{b})). \tag{8}$$

L'utilisation simultanée de (7) et (8) fait que la matrice se multiplie par le vecteur une seule fois par pas, à savoir lors du calcul de $A(Ax^n - b)$; quant aux vecteurs $Ax^{n+1} - b$, ils se déterminent par la relation de récurrence (8).

Cependant, dans cet algorithme, comme lors de l'utilisation directe de la formule (7), il y a une accumulation importante de l'erreur de calcul. On l'évite en modifiant l'algorithme. La formule (7) s'écrit

$$x^{n+1} = x^n + \alpha_n (x^n - x^{n-1}) + \beta_n (Ax^n - b), \tag{9}$$

où

$$\beta_n = \frac{-1}{\sigma_n^n + \sigma_{n-1}^n}, \quad \alpha_n = \beta_n \sigma_{n-1}^n.$$

Au lieu de trouver α_n et β_n par ces dernières formules, on les déterminera chaque fois en minimisant la fonctionnelle $F(x^{n+1})$ par rapport à α_n et β_n. Il vient les formules

$$\begin{aligned}
\alpha_n &= ((Ax^n - b,\ Ax^n - b)(Ax^n - b,\ A(x^n - x^{n-1})) - \\
&\quad - (Ax^n - b,\ x^n - x^{n-1})(A(Ax^n - b),\ Ax^n - b))\,\sigma_n, \\
\beta_n &= ((Ax^n - b,\ x^n - x^{n-1})(A(x^n - x^{n-1}),\ Ax^n - b) - \\
&\quad - (Ax^n - b,\ Ax^n - b)(x^n - x^{n-1},\ A(x^n - x^{n-1})))\,\sigma_n, \qquad (10) \\
\sigma_n &= \frac{1}{(x^n - x^{n-1},\ A(x^n - x^{n-1}))} \times \\
&\quad \times \frac{1}{(Ax^n - b,\ A(Ax^n - b)) - (Ax^n - b,\ A(x^n - x^{n-1}))^2}.
\end{aligned}$$

Chaque pas d'itération exige déjà trois opérations de multiplication de la matrice par le vecteur. En l'absence d'arrondis, ces processus itératifs fournissent évidemment une même suite de vecteurs x^n. Le dernier processus est cependant auto-correcteur: l'influence des erreurs d'arrondi commises aux pas précédents s'amortit progressivement. Tout comme pour la méthode de la plus grande pente, il serait peut-être raisonnable de s'adresser à tour de rôle aux formules (7)-(8) et (9)-(10). Si l'on n'arrondit pas, on a $(Ax^n - b,\ x^n - x^{n-1}) = 0$; pourtant, en remplaçant celle-ci par 0 dans (10), le processus risque de ne plus être auto-correcteur.

Estimons la vitesse de convergence. Pour les approximations y^n du processus linéaire optimal on a:

si

$$x^0 - X = y^0 - X = \sum_i c_i e_i,$$

alors

$$y^n - X = \sum_i c_i g_i e_i,$$

où

$$|g_i| \leqslant |t_n|^{-1} = \left| T_n\left(-\frac{M+\mu}{M-\mu}\right)\right|^{-1}.$$

D'où l'estimation

$$F_0(\mathbf{y}^n) = (A(\mathbf{y}^n - \mathbf{X}), \mathbf{y}^n - \mathbf{X}) = \\ = \sum_i \lambda_i c_i^2 g_i^2 \leqslant |t_n|^{-2} \sum_i \lambda_i c_i^2 = |t_n|^{-2} (A(\mathbf{x}^0 - \mathbf{X}), \mathbf{x}^0 - \mathbf{X}).$$

De toutes les approximations du type (1) dont $\mathbf{y}^n$, c'est l'approximation de la méthode du gradient conjugué qui minimise les fonctionnelles F et F_0. On a donc

$$F_0(\mathbf{x}^n) \leqslant F_0(\mathbf{y}^n) \leqslant |t_n|^{-2} F_0(\mathbf{x}^0).$$

Evaluons comme plus haut la vitesse de convergence par la norme usuelle

$$\| \mathbf{x}^n - \mathbf{X} \|_3 \leqslant \sqrt{\frac{M}{\mu}} |t_n|^{-1} \| \mathbf{x}^0 - \mathbf{X} \|_3.$$

Nous voyons que l'évaluation est du même ordre de grandeur que celle pour le processus linéaire optimal. Le grand avantage de la méthode utilisant les directions conjuguées est qu'elle néglige les bornes du spectre.

Puisque après $q \leqslant m$ itérations on aboutit, sous l'hypothèse de l'absence d'arrondis, à une solution exacte, cette méthode peut être qualifiée de directe. Soit une matrice A possédant l éléments non nuls avec naturellement $l \geqslant m$. Un pas d'itération exige $O(l)$ opérations arithmétiques et une solution rigoureuse ne s'obtient donc qu'au bout de $qO(l) = O(ml)$ opérations. Ainsi, pour les matrices renfermant un faible nombre d'éléments non nuls, notre procédé s'avère direct et efficace. De plus, on ne constate dans ce procédé qu'un encombrement relativement faible de mémoires des machines à calculer. Si l'ordre m du système est élevé, il n'est pas rare qu'une bonne approximation de la solution se réalise pour un nombre d'itérations $n \ll q$.

Considérons le cas où la condition $A > 0$ n'est pas réalisée. Le système $A^*A\mathbf{x} = A^*\mathbf{b}$ est équivalent au système initial, mais le nombre d'éléments non nuls de la matrice A^*A peut dépasser notablement l. Il serait alors raisonnable de ne pas expliciter ce système et de trouver chaque vecteur de la forme $A^*A\mathbf{u}$ en multipliant successivement $\mathbf{u}$ par A et A^*.

§ 11. Méthode de Monte-Carlo pour des systèmes d'équations linéaires

Dans les méthodes par itérations que nous venons d'appliquer à des systèmes d'équations linéaires, les mémoires des machines à calculer doivent conserver toutes les composantes de la solution, ce qui présente un grand inconvénient lorsque m est très important. Si l'ordre du système est fort élevé et les coefficients de celui-ci se calculent moyennant des formules élémentaires, on peut se servir de la méthode de Monte-Carlo.

Examinons le schéma général de la méthode. Proposons-nous de calculer la somme $S = \sum_{k=1}^{N} b_k$. Mettons chaque b_k sous la forme $b_k = p_k f_k$, où $p_k \geqslant 0$, $\sum_{k=1}^{N} p_k = 1$. Soit ξ une variable aléatoire prenant la valeur k avec la probabilité p_k. Choisissons M valeurs indépendantes $k_1, \ldots, k_M$ de cette variable et posons

$$S \approx S_M = \frac{1}{M} \sum_{l=1}^{M} f_{k_l}.$$

On a

$$\mathrm{E} f_k = \sum_{k=1}^{N} p_k f_k = \sum_{k=1}^{N} b_k = S, \tag{1}$$

$$d^2 = \mathrm{Var}\, f_k = \mathrm{E} f_k^2 - (\mathrm{E} f_k)^2 = \sum_{k=1}^{N} p_k f_k^2 - S^2 = \sum_{k=1}^{N} \frac{b_k^2}{p_k} - S^2.$$

Les valeurs f_k étant indépendantes, on a $\mathrm{Var}\,(S_M) = \frac{d^2}{M}$. L'inégalité de Tchebycheff nous donne l'estimation: pour $0 < \eta < 1$, la relation $|S_M - S| \leqslant d/\sqrt{\eta M}$ est valable avec la probabilité $1 - \eta$. Ainsi, si M est important, la probabilité pour S_M d'être proche de S est très grande.

On obtient les valeurs de ξ de la façon suivante. La machine est d'habitude dotée d'un dispositif générant des nombres aléatoires ou chargée d'un programme de nombres pseudo-aléatoires donnant les valeurs $\eta_1, \ldots$ d'une variable aléatoire uniformément répartie sur [0, 1).

On détermine pour chaque η_l un nombre k_l tel que η_l appartienne à l'intervalle semi-ouvert.

$$\pi_{k_l} = \Big[\sum_{i=1}^{k_l - 1} p_i,\ \sum_{i=1}^{k_l} p_i\Big)$$

et on pose $\xi_l = k_l$. La probabilité pour que η_l tombe sur l'intervalle semi-ouvert π_s est égale à p_s; donc la variable ξ ainsi obtenue prend chaque valeur s avec la probabilité p_s. Le plus simple serait de prendre $p_k \equiv \frac{1}{N}$, car alors $\sum_{i=1}^{l} p_i = \frac{1}{N}$ et ξ_l se déterminerait par la formule simple

$$\xi_l = [\eta_l N] + 1.$$

Cependant, à la lumière de (1) on voit que la quantité d qui caractérise l'exactitude du résultat dépend essentiellement du choix de p_k. Un tel algorithme simple laisse très souvent à désirer et la résolution de problèmes concrets par la méthode de Monte-Carlo exige un gros effort quant au choix optimal des p_k pour un ensemble donné des nombres b_k.

Ci-dessous un cas important lorsque tout les $b_k \geqslant 0$. Minimisons d^2 par la méthode de Lagrange dans la condition $\sum_{k=1}^{N} p_k = 1$. La fonction de Lagrange

s'écrit

$$\Phi(p_1, \ldots, p_N) = \sum_{k=1}^{N} \frac{b_k^2}{p_k} - S^2 + \lambda \left(\sum_{k=1}^{N} p_k - 1 \right).$$

La condition $\Phi'_{p_k} = 0$, $k = 1, \ldots, N$, implique $-\frac{b_k^2}{p_k^2} + \lambda = 0$. D'où $p_k = \frac{b_k}{\sqrt{\lambda}}$. La condition $\sum_{k=1}^{N} p_k = 1$ entraîne $\sum_{k=1}^{N} b_k = \sqrt{\lambda}$, i.e. $\sqrt{\lambda} = S$. La valeur $p_k = \frac{b_k}{S}$ est, on le voit, la plus indiquée. Pour ce choix

$$f_k = \frac{b_k}{p_k} \equiv S, \quad d^2 = 0.$$

Pour distribuer optimalement p_k il faut connaître la quantité S cherchée. Voici un des modes opératoires possibles: on choisit une fonction $B_k \geqslant 0$ telle que $b_k \approx B_k$ et les sommes $\sigma_n = \sum_{k=1}^{L} B_k$ soient connues quel que soit n; ensuite on pose $p_k = \frac{B_k}{\sigma_N}$.

Appliquons la méthode de Monte-Carlo à la résolution de systèmes d'équations linéaires. Supposons qu'on cherche la composante $(\mathbf{x})_q$ de la solution du système $\mathbf{x} = B\mathbf{x} + \mathbf{c}$. Soit B une matrice telle que le processus itératif $\mathbf{x}^{n+1} = B\mathbf{x}^n + \mathbf{c}$ converge vers la solution $\mathbf{X}$ du système pour toute approximation initiale $\mathbf{x}^0$. Lorsque $\mathbf{x}^0 = 0$, il vient

$$\mathbf{x}^1 = \mathbf{c}, \quad \mathbf{x}^2 = \mathbf{c} + B\mathbf{c}, \ldots, \mathbf{x}^{n+1} = \sum_{j=0}^{n} B^j \mathbf{c}.$$

Le processus étant convergent, pour n suffisamment élevé, la quantité $(\mathbf{x}^{n+1})_q$ approche assez bien $(\mathbf{X})_q$. La proximité de $\mathbf{x}^{n+1}$ et $\mathbf{X}$ nous guide dans le choix de n. Cherchons ensuite $(\mathbf{x}^{n+1})_q$. On a

$$(x^{n+1})_q = c_q + \sum_{j_1=1}^{m} b_{qj_1} c_{j_1} + \ldots + \sum_{j_1=1}^{m} \ldots \sum_{j_m=1}^{m} b_{qj_1} \ldots b_{j_{n-1}j_n} c_{j_n}. \tag{2}$$

Ainsi, sur le plan formel, rechercher $(\mathbf{x}^{n+1})_q$ revient à rechercher S. Cependant, lorsque N sont très importants, il est inutile d'utiliser le procédé examiné plus haut en raison de la longueur limitée des nombres avec lesquels travaille la machine. En effet, les nombres sont arrondis à des quantités de la forme $k2^{-t}$, k entier, t la longueur des nombres traités. Aussi les longueurs des intervalles semi-ouverts π_l obtenus dans la pratique seront multiples de 2^{-t} et, dans le cas des p_l comparables avec 2^{-t}, l'erreur sur les valeurs de la fonction aléatoire ξ sera considérable. En particulier, cette longueur limitée fait que le nombre d'intervalles semi-ouverts non nuls π_l est toujours 2^t au plus. La formule (2) comprend $N > m^n$ termes et le procédé examiné est inapplicable directement même pour m et n pas très forts.

Essayons de surmonter cette difficulté. Soit $p_{ij} \geqslant 0$ et vérifient

$$\sum_{j=1}^{m} p_{ij} = 1 \text{ pour } i = 1, \ldots, m, \; p_{ij} > 0 \text{ si } b_{ij} \neq 0. \tag{3}$$

Soient encore ξ^i, $i = 1, \ldots, m$ des variables aléatoires indépendantes telles que ξ^i prend la valeur j avec la probabilité p_{ij}. Tirons au sort la valeur de la variable aléatoire ξ^q, soit j_1 la valeur ainsi obtenue. Faisons de même pour la variable ξ^{j_1} et obtenons, disons, j_2, et ainsi de suite; choisissons au hasard la valeur de la variable aléatoire $\xi^{j_{n-1}}$, soit j_n cette valeur. Posons

$$\overline{S}_{j_1, \ldots, j_n} = c_q + f_{qj_1} c_{j_1} + \ldots + f_{qj_1} \cdots f_{j_{n-1}j_n} c_{j_n},$$

où $f_{ij} = b_{ij}/p_{ij}$ et $f_{ij} = 0$ si $b_{ij} = 0$, $p_{ij} = 0$. En effectuant cet échantillonnage M fois, on aboutit aux M suites $j_1^l, \ldots, j_n^l$ et aux M valeurs $\overline{S}_{j_1^l, \ldots, j_n^l}$.

Faisons

$$(x^{n+1})_q \approx \overline{S}^{(M)} = \frac{1}{M} \sum_{l=1}^{M} \overline{S}_{j_1^l, \ldots, j_n^l}.$$

Calculons $E\overline{S}_{j_1, \ldots, j_n}$. La probabilité d'obtenir la suite $j_1, \ldots, j_n$ vaut $p_{qj_1} \cdots p_{j_{n-1}j_n}$. Donc,

$$E\overline{S}_{j_1, \ldots, j_n} = \sum_{j_1=1}^{m} \cdots \sum_{j_n=1}^{m} (c_q + f_{qj_1} c_{j_1} + \ldots + f_{qj_1} \cdots f_{j_{n-1}j_n} c_{j_n}) p_{qj_1} \cdots p_{j_{n-1}j_n}.$$

Représentons le deuxième membre de cette égalité par $\sum_{k=0}^{n} T_k$ avec

$$T_0 = \sum_{j_1=1}^{m} \cdots \sum_{j_n=1}^{m} c_q p_{qj_1} \cdots p_{j_{n-1}j_n},$$

$$T_k = \sum_{j_1=1}^{m} \cdots \sum_{j_n=1}^{m} f_{qj_1} \cdots f_{j_{k-1}j_k} c_{j_k} p_{qj_1} \cdots p_{j_{n-1}j_n} \quad \text{pour } k > 0.$$

Soit

$$\sigma_j^{k, n} = \sum_{j_{k+1}=1}^{m} \cdots \sum_{j_n=1}^{m} p_{jj_{k+1}} \cdots p_{j_{n-1}j_n},$$

alors

$$T_0 = c_q \sigma_q^{0, n},$$

$$T_k = \sum_{j_1=1}^{m} \cdots \sum_{j_k=1}^{m} (f_{qj_1} p_{qj_1}) \cdots (f_{j_{k-1}j_k} p_{j_{k-1}j_k}) c_{j_k} \sigma_{j_k}^{k, n}.$$

Une sommation successive sur les variables $j_n, \ldots, j_{k+1}$ à l'aide de (3) fournit $\sigma_j^{k, n} = 1$. Il s'ensuit, compte tenu des égalités $f_{ij} p_{ij} = b_{ij}$, que

$$T_0 = c_q,$$

$$T_k = \sum_{j_1=1}^{m} \cdots \sum_{j_k=1}^{m} b_{qj_1} \cdots b_{j_{k-1}j_k} c_{j_k},$$

$k > 0$. Ainsi, $T_k = (B^k \mathbf{c})_q$; compte tenu des égalités

$$\mathrm{E}\bar{S}_{j_1, \ldots, j_n} = \sum_{k=0}^{n} T_k$$

et de (2) il vient

$$\mathrm{E}\bar{S}_{j_1, \ldots, j_n} = (\mathbf{x}^{n+1})\, q.$$

D'où

$$\mathrm{E}\bar{S}^{(M)} = (\mathbf{x}^{n+1})\, q.$$

Puisque $\mathrm{Var}\,(\bar{S}^{(M)}) = \dfrac{\mathrm{Var}\,\bar{S}_{j_1, \ldots, j_n}}{M}$, pour M élevés la quantité $\bar{S}^{(M)}$ est voisine avec une grande probabilité de $(\mathbf{x}^{n+1})_q$ et on peut estimer que $(\mathrm{X})_q \approx \approx \bar{S}^{(M)}$. Les valeurs $\bar{S}_{j_1, \ldots, j_n}$ se calculent lors du tirage au sort des $j_1, \ldots, j_n$ par les formules de récurrence

$$\begin{cases} \bar{S}_{j_1, \ldots, j_l} = \bar{S}_{j_1, \ldots, j_{l-1}} + z_{j_1, \ldots, j_l} c_{j_l}, \\ z_{j_1, \ldots, j_l} = z_{j_1, \ldots, j_{l-1}} f_{j_{l-1}, j_l} \text{ pour } l > 1, \end{cases}$$

avec

$$\bar{S}_{j_1} = c_q + z_{j_1} c_{j_1}, \qquad z_{j_1} = f_{qj_1}.$$

Si l'on réalise l'algorithme moyennant ces formules, le calcul de $\bar{S}^{(M)}$ exige la mise en réserve d'un ensemble fini (non croissant lorsque $M, n \to \infty$) de nombres.

De même que dans le cas d'intégrales de multiplicité élevée, la résolution des systèmes d'équations linéaires par la technique de Monte-Carlo exige beaucoup de prudence. En particulier, il convient de choisir avec le plus grand soin l'ensemble des p_{ij}.

Illustrons nos paroles à l'aide d'un exemple modèle élémentaire. Soit

$$x'' (t) = 0, \; x (0) = c_0, \; x (1) = c_1 \tag{4}$$

une équation à résoudre sur le segment $0 \leqslant t \leqslant 1$. Donnons-nous le pas $h = 1/L$, $L = 2l$ (l entier) et écrivons l'approximation discrète

$$x_0 = c_0, \; x_{k+1} - 2x_k + x_{k-1} = 0, \; k = 1, \ldots, L-1, \; x_L = c_1 \tag{5}$$

(dans notre cas le problème initial (4) est approximé sans erreur). Mettons (5) sous la forme

$$\begin{aligned} x_0 &= c_0, \\ x_k &= \frac{x_{k-1} + x_{k+1}}{2}, \qquad k = 1, \ldots, L-1, \\ x_L &= c_1. \end{aligned} \tag{6}$$

Soit $\mathbf{x} = B\mathbf{x} + \mathbf{c}$ l'écriture matricielle des relations (6). Cherchons la composante x_l. La matrice B est symétrique et

$$\lambda_B^i = \cos \frac{\pi i}{L}, \qquad i = 1, \ldots, L-1.$$

Aussi

$$\max |\lambda_B^i| = \cos\frac{\pi}{L},$$

$$\| B^n \|_3 = \left(\cos\frac{\pi}{L}\right)^n = \exp\left(n \ln\cos\frac{\pi}{L}\right) = \exp\left(-n\left(\frac{\pi^2}{2L^2} + O\left(\frac{\pi}{L}\right)^4\right)\right). \tag{7}$$

La condition $\| B^n \|_3 \ll 1$ est essentielle si l'on veut que l'erreur sur la valeur approchée $(x^{n+1})_l$ soit petite. On prend donc

$$n \gg L^2. \tag{8}$$

Supposons, par exemple, que tous les $p_{ij} = \frac{1}{L+1}$ et, pour fixer les idées, que n est pair. Alors

$$f_{ij} = \begin{cases} \frac{L+1}{2} & \text{pour } |i-j| = 1 \text{ si } (i, j) \neq (0, 1), (L, L-1), \\ 0 & \text{pour les } (i, j) \text{ restants.} \end{cases}$$

Si $n > l$, la quantité $\bar{S}_{j_1, \ldots, j_n}$ peut renfermer le terme

$$f_{lj_1} f_{j_1 j_2} \cdots f_{j_{n-1} j_n} c_0,$$

$|j_{p+1} - j_p| = 1$ avec $p = 1, \ldots, n-1$. Ce terme vaut $\left(\frac{L+1}{2}\right)^n c_0$ et la probabilité de son apparition est $(L+1)^{-n}$. Si ξ est une variable aléatoire prenant les valeurs y_l avec les probabilités q_l, alors

$$\operatorname{Var}\xi = \sum_i q_i (y_i - \mathrm{E}\xi)^2 \geqslant q_l (y_l - \mathrm{E}\xi)^2$$

avec l quelconque. C'est pourquoi

$$\operatorname{Var}\bar{S}_{j_1, \ldots, j_n} \geqslant (L+1)^{-n} \left(\left(\frac{L+1}{2}\right)^n c_0 - \mathrm{E}\bar{S}_{j_1, \ldots, j_n}\right)^2.$$

Nous avons choisi un n tel que $\mathrm{E}\bar{S}_{j_1, \ldots, j_n} = (x^{n+1})_l \approx x_l$. Du moment que $x_l = (c_0 + c_1)/2$, le terme principal dans le second membre de cette inégalité est de l'ordre de $((L+1)/4)^n$. La comparaison avec (8) révèle que déjà pour L peu importants on n'a guère d'espoir de calculer x_l avec une précision acceptable pour un M (nombre d'épreuves) raisonnable.

Soit maintenant

$$p_{ij} = \begin{cases} \frac{1}{2} & \text{pour } i \neq 0, L, \ |i-j| = 1, \\ 1 & i = j = 0, L, \\ 0 & \text{dans tous les cas restants.} \end{cases} \tag{9}$$

Considérons une particule errant d'un point jh $(j = 0, \ldots, L)$ à l'autre. Supposons que ce mouvement commence à l'instant $\tau = 0$ au point lh et que de chaque point jh, $j \neq 0, L$, la particule passe par unité de temps à l'un

des points $(j \pm 1)\,h$ et ce avec la probabilité 1/2; si la particule tombe en un des points $t = 0$, $t = 1$, elle y reste. Soient ih, $j_1 h$, . . ., $j_n h$ les positions successives de la particule; posons

$$\widetilde{S}_{j_1, \ldots, j_n} = \begin{cases} c_0 & \text{pour} \quad j_n = 0, \\ 0 & \text{pour} \quad j_n \neq 0, L, \\ c_1 & \text{pour} \quad j_n = L. \end{cases}$$

On vérifie aisément que la probabilité de cette trajectoire coïncide avec la probabilité d'apparition de la suite $j_1, \ldots, j_n$ dans la méthode de Monte-Carlo de résolution du système initial avec la table de probabilités (9) et $\widetilde{S}_{j_1, \ldots, j_n} = \overline{S}_{j_1, \ldots, j_n}$. Dans notre cas

$$|\overline{S}_{j_1, \ldots, j_n}| \leqslant \overline{c} = \max(|c_0|, |c_1|)$$

et, par conséquent, $\operatorname{Var} \overline{S}_{j_1, \ldots, j_n} \leqslant \overline{c}^2$ quels que soient L et n.

Ceci nous permet d'espérer une solution d'une certaine précision pour un coût acceptable en temps machine. En appréciant la qualité de la méthode décrite on n'oubliera pas que, vu (8), le temps d'échantillonnage d'un seul ensemble $j_1, \ldots, j_n$ augmente avec L. Aussi, lorsqu'il s'agit d'approcher les problèmes aux limites pour des équations différentielles ordinaires, la méthode de Monte-Carlo ne peut concurrencer les méthodes directes.

Une autre remarque de caractère général. Soit q_n la probabilité de l'événement $f_{ij_1} \ldots f_{j_{n-1}j_n} = 0$. Alors $q_n - q_{n-1}$ est égal à la probabilité de l'événement $f_{ij_1} \ldots f_{j_{n-2}j_{n-1}} \neq 0$ et $f_{ij_1} \ldots f_{j_{n-1}j_n} = 0$. Si q_n tend vers l'unité pour $n \to \infty$, la quantité

$$Q = \sum_{n=1}^{\infty} (q_n - q_{n-1})\, n$$

est égale à l'espérance mathématique de la valeur n pour laquelle le produit $f_{ij_1} \ldots f_{j_{n-1}j_n}$ s'annule une première fois.

Dans les problèmes appliqués on arrive souvent à choisir p_{ij} de façon que les variances $\operatorname{Var} S_{j_1, \ldots, j_n}$ soient uniformément bornées en n et que Q ne soit pas trop élevé. Dans ce cas, on ne fixe pas d'avance le nombre n et on choisit au hasard les éléments de l'ensemble $j_1, \ldots, j_n$ tant qu'on n'aura pas

$$f_{j_{n-1}j_n} = 0.$$

PROBLÈME. Démontrer que pour un tel algorithme

$$\mathrm{E}S_{j_1, \ldots\, j_n} = (\mathbf{x})_q.$$

Dans la méthode de Monte-Carlo, il importe également de faire attention à la représentation du système sous la forme $\mathbf{x} = B\mathbf{x} + \mathbf{c}$ pour laquelle le processus itératif $\mathbf{x}^{n+1} = B\mathbf{x}^n + \mathbf{c}$ converge le plus rapidement possible (cela fait décroître le nombre n, i.e. allège une épreuve). A défaut d'un choix heureux des p_{ij}, une bonne approximation $\mathbf{x}^0$ de la solution entraîne aussi une diminution de $\operatorname{Var} \overline{S}_{j_1, \ldots, j_n}$. Dans ce cas, on résout le système

$$\mathbf{r} = B\mathbf{r} + (\mathbf{c} + B\mathbf{x}^0 - \mathbf{x}^0)$$

relativement à $\mathbf{r} = \mathbf{x} - \mathbf{x}^0$ par la technique de Monte-Carlo; il faut s'attendre alors à une diminution de $\operatorname{Var} \overline{S}_{j_1, \ldots, j_n}$ et donc à une réduction du volume de calcul.

§ 12. Méthode par itérations avec utilisation des opérateurs spectralement équivalents

A part des méthodes itératives simples telles que

$$\mathbf{x}^{n+1} = \mathbf{x}^n - \alpha (A\mathbf{x}^n - \mathbf{b}) \tag{1}$$

on en applique souvent d'autres, à savoir

$$B\mathbf{x}^{n+1} = B\mathbf{x}^n - \alpha (A\mathbf{x}^n - \mathbf{b}), \tag{2}$$

$B \neq E$ étant une matrice facilement inversible. La multiplication de (2) par B^{-1} donne

$$\mathbf{x}^{n+1} = \mathbf{x}^n - \alpha B^{-1} (A\mathbf{x}^n - \mathbf{b}),$$

et le processus itératif (2) est, on le voit, équivalent à la méthode itérative simple de matrice $E - \alpha B^{-1}A$.

Considérons le cas où $A > 0$ et le rapport de la plus grande valeur propre M à la plus petite μ de la matrice A est important. Les processus par itérations que nous avons examinés plus haut convergent alors lentement. Soit $B > 0$ et

$$\mathrm{M}_1 = \sup_{\mathbf{x}\neq 0} \frac{(A\mathbf{x}, \mathbf{x})}{(B\mathbf{x}, \mathbf{x})}, \quad \mu_1 = \inf_{\mathbf{x}\neq 0} \frac{(A\mathbf{x}, \mathbf{x})}{(B\mathbf{x}, \mathbf{x})}. \tag{3}$$

Supposons que B s'inverse facilement et $\mathrm{M}_1/\mu_1 \ll \mathrm{M}/\mu$.

Dans le cas de M_1/μ_1 pas très grand, on dit des méthodes itératives que nous allons examiner dans le présent paragraphe qu'elles utilisent des *opérateurs spectralement équivalents*. Montrons que pour un α bien choisi, la méthode (2) converge plus rapidement que le processus élémentaire (1).

La solution X exacte vérifie l'égalité

$$B\mathrm{X} = B\mathrm{X} - \alpha (A\mathrm{X} - \mathbf{b}),$$

en la retranchant de (2) il vient une équation relativement à l'erreur

$$B\mathbf{r}^{n+1} = B\mathbf{r}^n - \alpha A\mathbf{r}^n. \tag{4}$$

Diagonalisons la matrice B par une transformation orthogonale. Soit $B = U^*\Lambda U$, où

$$\Lambda = \begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_m \end{pmatrix}$$

et U une matrice orthogonale.

On désigne habituellement par $\sqrt{B}$ la matrice

$$U^* \begin{pmatrix} +\sqrt{\lambda_1} & & 0 \\ & \ddots & \\ 0 & & +\sqrt{\lambda_m} \end{pmatrix} U.$$

Evidemment, $\sqrt{B} > 0$, $(\sqrt{B})^* = \sqrt{B}$, $\sqrt{B}\sqrt{B} = B$. Multiplions à gauche par $(\sqrt{B})^{-1}$ les deux membres de (4) et introduisons les nouvelles quantités $\sqrt{B}\mathbf{r}^n = \mathbf{v}^n$. Il vient

$$\mathbf{v}^{n+1} = \mathbf{v}^n - \alpha C \mathbf{v}^n \tag{5}$$

avec $C = (\sqrt{B})^{-1} A (\sqrt{B})^{-1}$.

Etant donné $A = A^*$, $(\sqrt{B}) = (\sqrt{B})^*$, la matrice C est symétrique. Examinons l'expression

$$w(\mathbf{x}) = \frac{(C\mathbf{x}, \mathbf{x})}{(\mathbf{x}, \mathbf{x})}.$$

Après avoir posé $(\sqrt{B})^{-1}\mathbf{x} = \mathbf{y}$, on peut écrire

$$w(\mathbf{x}) = w(\sqrt{B}\,\mathbf{y}) = \frac{(A\mathbf{y}, \mathbf{y})}{(B\mathbf{y}, \mathbf{y})}.$$

Par suite de (3), on a $w(\mathbf{x}) \in [\mu_1, M_1]$; c'est pourquoi toutes les valeurs propres de C appartiennent également au segment $[\mu_1, M_1]$.

Pour $\alpha = \frac{2}{M_1 + \mu_1}$ les valeurs propres de la matrice $E - \alpha C$ ont un module égal à $\frac{M_1 - \mu_1}{M_1 + \mu_1}$ au plus. Aussi $\| E - \alpha C \|_3 \leqslant \frac{M_1 - \mu_1}{M_1 + \mu_1}$ et de façon analogue à (7.7) on a

$$\| \mathbf{v}^n \|_3 \leqslant \left(\frac{M_1 - \mu_1}{M_1 + \mu_1}\right)^n \| \mathbf{v}^0 \|_3. \tag{6}$$

On rappelle que

$$\| \mathbf{v}^n \|_3 = \sqrt{(\sqrt{B}\,\mathbf{r}^n, \sqrt{B}\,\mathbf{r}^n)} = \sqrt{(B\mathbf{r}^n, \mathbf{r}^n)}.$$

Si $0 < \mu_2 \leqslant \frac{(B\mathbf{y}, \mathbf{y})}{(\mathbf{y}, \mathbf{y})} \leqslant M_2$ pour n'importe quel y, (6) entraîne l'estimation

$$\| \mathbf{r}^n \|_3 \leqslant \frac{\| \mathbf{v}^n \|_3}{\sqrt{\mu_2}} \leqslant \left(\frac{M_1 - \mu_1}{M_1 + \mu_1}\right)^n \frac{\| \mathbf{v}^0 \|_3}{\sqrt{\mu_2}} \leqslant \left(\frac{M_1 - \mu_1}{M_1 + \mu_1}\right)^n \sqrt{\frac{M_2}{\mu_2}}\, \| \mathbf{r}^0 \|_3.$$

Puisque la fonction

$$\frac{M - \mu}{M + \mu} = \frac{1 - (\mu/M)}{1 + (\mu/M)}$$

diminue de façon monotone avec M/μ, le nouveau processus itératif converge plus rapidement.

Avant de passer des formules (1) aux formules (2), il convient de comparer l'effort qu'elles exigent du chercheur. Si les itérations par les formules (2) sont sensiblement plus laborieuses, ce passage est peut être à déconseiller.

PROBLÈME. Soient $T(A)$ le travail qu'exige une itération (1) et $T(B)$ celui dans le cas de la formule (2). Justifier le passage aux itérations (2) si

$$\frac{T(B)}{\ln\frac{M_1+\mu_1}{M_1-\mu_1}} < \frac{T(A)}{\ln\frac{M+\mu}{M-\mu}}.$$

En s'inspirant de la façon dont la méthode (2) a été construite à partir de la méthode (1), on élabore un processus itératif correspondant à chacun des procédés par itérations considérés aux §§ 7, 9, 10.

Soit un processus itératif tel que les erreurs sur les approximations successives vérifient l'égalité

$$\mathbf{r}^{n+1} = \left(E - \sum_{i=1}^{k} \alpha_i A^i\right) \mathbf{r}^n.$$

Ecrivons

$$\mathbf{v}^{n+1} = \left(E - \sum_{i=1}^{k} \alpha_i C^i\right) \mathbf{v}^n \tag{7}$$

avec $C = (\sqrt{B})^{-1} A (\sqrt{B})^{-1}$, posons $\mathbf{v}^n = \sqrt{B}\,\mathbf{r}_n$ et multiplions à gauche (7) par $\sqrt{B}$. Il vient

$$B\mathbf{r}^{n+1} = B\mathbf{r}^n - \sum_{i=1}^{k} \alpha_i (AB^{-1})^{i-1} A\mathbf{r}^n.$$

Cette équation d'erreur correspond au processus itératif

$$B\mathbf{x}^{n+1} = B\mathbf{x}^n - \sum_{i=1}^{k} \alpha_i (AB^{-1})^{i-1} (A\mathbf{x}^n - \mathbf{b}). \tag{8}$$

Par une optimisation des paramètres α_i telle que l'évaluation du rapport des normes $\|\mathbf{v}^{n+1}\|_3/\|\mathbf{v}^n\|_3$ soit la meilleure, on aboutit, tout comme au § 7, à la meilleure méthode itérative de la forme (8).

L'analogue de la méthode linéaire optimale (7.16), (7.17) est le processus itératif

$$B\mathbf{y}^{n+1} = B\mathbf{y}^n + \bar{\omega}_n \bar{\omega}_{n-1} B(\mathbf{y}^n - \mathbf{y}^{n-1}) - \frac{2}{M_1+\mu_1}(1 + \bar{\omega}_n \bar{\omega}_{n-1})(A\mathbf{y}^n - \mathbf{b})$$

avec $\bar{\omega}_n$ construit à l'aide de M_1 et μ_1 par analogie avec la construction de ω_n d'après M et μ.

Voyons une technique qui combine le processus (2) avec la méthode de la plus grande pente. L'approximation $\mathbf{x}^{n+1}$ s'obtient moyennant la relation

$$B\mathbf{x}^{n+1} = B\mathbf{x}^n - \alpha_n (A\mathbf{x}^n - \mathbf{b}), \tag{9}$$

le coefficient α_n étant déterminé par la condition de minimum

$$F(\mathbf{x}) = (A\mathbf{x}, \mathbf{x}) - 2(\mathbf{b}, \mathbf{x});$$

(9) fournit

$$B\frac{d\mathbf{x}^{n+1}}{d\alpha_n} = -(A\mathbf{x}^n - \mathbf{b}).$$

Comme

$$\frac{dF(\mathbf{x}^{n+1})}{d\alpha_n}=2\left(A\mathbf{x}^{n+1}-\mathbf{b},\ \frac{d\mathbf{x}^{n+1}}{d\alpha_n}\right)=2\,(A\,(\mathbf{x}^n-\alpha_n B^{-1}\,(A\mathbf{x}^n-\mathbf{b}))-$$
$$-\mathbf{b},\ -B^{-1}\,(A\mathbf{x}^n-\mathbf{b})),$$

la condition $\frac{dF(\mathbf{x}^{n+1})}{d\alpha_n}=0$ est une équation linéaire en α_n.

L'analogue du processus itératif (10.9), (10.10) s'écrit

$$B\mathbf{x}^{n+1}=B\mathbf{x}^n+\alpha_n\,(B\mathbf{x}^n-B\mathbf{x}^{n-1})+\beta_n\,(A\mathbf{x}^n-\mathbf{b}).$$

Les méthodes examinées dans le présent paragraphe sont d'un usage fréquent lorsqu'il faut résoudre les systèmes d'équations $A\mathbf{x}=\mathbf{b}$ où les valeurs propres de la matrice $A>0$ sont très dispersées.

§ 13. Erreur sur la solution approchée d'un système d'équations et conditionnement des matrices. Régularisation

Supposons que la matrice et le deuxième membre du système ne sont donnés qu'avec une certaine précision et qu'au lieu du système proposé

$$A\mathbf{x}=\mathbf{b}, \tag{1}$$

c'est

$$A_1\mathbf{x}=\mathbf{b}_1,\ A_1=A+\Delta,\ \mathbf{b}_1=\mathbf{b}+\eta \tag{2}$$

qu'on doit résoudre. Supposons connues les estimations $\|\Delta\|$ et $\|\eta\|$. Evaluons l'erreur sur la solution.

Dégageons pour commencer la partie principale de l'erreur. Notons $\mathbf{X}$ et $\mathbf{X}^*$ les solutions respectives de (1) et (2) et $\mathbf{r}$ la différence $\mathbf{X}^*-\mathbf{X}$. Portons les expressions de A_1, $\mathbf{b}_1$ et $\mathbf{X}^*$ dans (2), il vient

$$(A+\Delta)\,(\mathbf{X}+\mathbf{r})=\mathbf{b}+\eta.$$

La soustraction de (1) donne

$$A\mathbf{r}+\Delta\mathbf{X}+\Delta\mathbf{r}=\eta,$$

d'où

$$A\mathbf{r}=\eta-\Delta\mathbf{X}-\Delta\mathbf{r}$$

et

$$\mathbf{r}=A^{-1}\,(\eta-\Delta\mathbf{X}-\Delta\mathbf{r}). \tag{3}$$

Si $\|\Delta\|$ et $\|\eta\|$ sont faibles, $\|\mathbf{r}\|$ doit l'être aussi. Le terme $\Delta\mathbf{r}$ est alors d'un ordre infinitésimal supérieur; en le négligeant on trouve

$$\mathbf{r}\approx A^{-1}\,(\eta-\Delta\mathbf{X}).$$

L'estimation d'erreur s'écrit donc

$$\|\mathbf{r}\|\leqslant\sigma\approx\|A^{-1}\|\,(\|\eta\|+\|\Delta\|\|\mathbf{X}\|). \tag{4}$$

On obtient une estimation rigoureuse de la façon suivante.

Par suite de (3) on a l'inégalité

$$\| \mathbf{r} \| \leqslant \| A^{-1} \| \| \eta \| + \| A^{-1} \| \| \Delta \| \| \mathbf{X} \| + \| A^{-1} \| \| \Delta \| \| \mathbf{r} \|.$$

Supposons que $\| A^{-1} \| \| \Delta \| < 1$. Portons le dernier terme de la somme dans le premier membre et divisons l'inégalité par le coefficient de $\| \mathbf{r} \|$. Il en résulte

$$\| \mathbf{r} \| \leqslant \frac{\| A^{-1} \| (\| \eta \| + \| \Delta \| \| \mathbf{X} \|)}{1 - \| A^{-1} \| \| \Delta \|}. \tag{5}$$

Il arrive souvent que l'erreur sur la matrice du système est sensiblement inférieure à celle due au deuxième membre; prenons comme modèle le cas où la matrice est donnée de façon exacte. En posant alors $\Delta = 0$ dans (5), il vient

$$\| \mathbf{r} \| \leqslant \| A^{-1} \| \| \eta \|.$$

Afin de caractériser qualitativement la relation entre l'erreur sur le membre à droite et celle sur la solution on fait appel aux notions de *conditionnement du système* et de *conditionnement de la matrice du système*. Les erreurs absolues résultant du deuxième membre et de la solution dépendent de l'échelle de ces quantités et de la matrice du système. Aussi est-il plus juste de caractériser les propriétés de celui-ci à l'aide de la relation qui lie les erreurs relatives correspondantes.

On mesure donc le *conditionnement du système* par le nombre

$$\tau = \sup_{\eta} \left(\frac{\| \mathbf{r} \|}{\| \mathbf{X} \|} : \frac{\| \eta \|}{\| \mathbf{b} \|} \right) = \frac{\| \mathbf{b} \|}{\| \mathbf{X} \|} \sup_{\eta} \frac{\| \mathbf{r} \|}{\| \eta \|}.$$

Conformément à cette définition l'erreur relative sur la solution est évaluée en fonction du nombre de conditionnement et de l'erreur relative sur le deuxième membre:

$$\frac{\| \mathbf{r} \|}{\| \mathbf{X} \|} \leqslant \tau \frac{\| \eta \|}{\| \mathbf{b} \|}. \tag{6}$$

Comme $\mathbf{r} = A^{-1}\eta$,

$$\sup_{\eta} \frac{\| \mathbf{r} \|}{\| \eta \|} = \| A^{-1} \|$$

et

$$\tau = \frac{\| \mathbf{b} \|}{\| \mathbf{X} \|} \| A^{-1} \|.$$

Parfois on a besoin d'une caractéristique plus grossière des propriétés du système, qui ne tienne compte que des propriétés de la matrice A. Cette caractéristique $\nu(A) = \sup_{\mathbf{b}} \tau$ s'appelle le *nombre de conditionnement de la matrice* A. Selon cette définition et (6) on a l'estimation

$$\frac{\| \mathbf{r} \|}{\| \mathbf{X} \|} \leqslant \nu(A) \frac{\| \eta \|}{\| \mathbf{b} \|}$$

qui relie les erreurs relatives sur le deuxième membre et sur la solution par l'intermédiaire des seules propriétés de la matrice du système. Puisque

$$\sup_{b} \frac{\| b \|}{\| X \|} = \sup_{x} \frac{\| Ax \|}{\| x \|} = \| A \|,$$

on a

$$\nu(A) = \| A \| \, \| A^{-1} \|.$$

Toute norme de la matrice est au moins égale à sa valeur propre de module le plus grand, i. e. $\| A \| \geqslant \max | \lambda_A |$. Les valeurs propres des matrices A et A^{-1} sont réciproquement inverses ; c'est pourquoi

$$\| A^{-1} \| \geqslant \max \frac{1}{| \lambda_A |} = \frac{1}{\min | \lambda_A |}.$$

Donc

$$\nu(A) \geqslant \frac{\max | \lambda_A |}{\min | \lambda_A |} \geqslant 1.$$

En particulier, si $A = A^*$, on a $\| A \|_3 = \max | \lambda_A |$, et

$$\| A^{-1} \|_3 = \max \frac{1}{| \lambda_A |} = \frac{1}{\min | \lambda_A |},$$

et alors

$$\nu(A) = \frac{\max | \lambda_A |}{\min | \lambda_A |}.$$

Etudions l'erreur commise sur la solution par suite de l'arrondissement du deuxième membre en calcul automatique. Soit, comme d'ordinaire, t la longueur des nombres binaires de la machine à calculer. Chaque élément b_i du deuxième membre est arrondi avec une erreur relative $O(2^{-t})$, c'est-à-dire avec une erreur absolue $O(|b_i| 2^{-t})$, aussi

$$\| \eta \| = O(\| b \| 2^{-t}) \quad \text{et} \quad \frac{\| \eta \|}{\| b \|} = O(2^{-t}).$$

Par conséquent

$$\frac{\| r \|}{\| X \|} \leqslant \nu(A) \, O(2^{-t}).$$

Dans la pratique, il est rare qu'on ait à évaluer strictement l'erreur sur la solution approchée d'un système d'équations linéaires à l'aide d'inégalités obtenues ou par un autre procédé. Toutefois l'information sur l'ordre de grandeur de cette erreur est souvent utile dans la mesure où elle permet d'établir des limites raisonnables à la précision des calculs. Les relations (4), (5) majorent l'erreur sur la solution résultant des erreurs de données. L'égalité (3) montre que

les estimations (4) et (5) sont assez exactes, et la recherche d'une solution affectée d'une erreur sensiblement inférieure à σ n'a pas de sens.

On dit des systèmes d'équations et des matrices qu'ils sont *mal conditionnés* si le nombre de conditionnement est élevé et qu'ils sont *bien conditionnés* dans le cas contraire. Si le deuxième membre de (4) qui évalue l'erreur sur la solution par les erreurs de données ou l'estimation de l'erreur de calcul sont grandes outre mesure, on aura intérêt à utiliser une information supplémentaire sur la solution du problème considéré. On abordera ce problème comme tout autre problème *mal posé*.

Examinons un cas simple de A symétrique. Soient $\lambda_1, \ldots, \lambda_m \neq 0$ ses valeurs propres ordonnées dans l'ordre de décroissance de $|\lambda_i|$ et $\mathbf{e}_1, \ldots, \mathbf{e}_m$ le système orthonormé correspondant de vecteurs propres. La solution du système $A\mathbf{x} = \mathbf{b}$ est le vecteur

$$\mathbf{X} = \sum_{i=1}^{m} c_i \mathbf{e}_i, \quad c_i = \frac{(\mathbf{b}, \mathbf{e}_i)}{\lambda_i}.$$

Pour $\tilde{\mathbf{b}} = \mathbf{b} + \boldsymbol{\eta}$ donné réellement, la solution est

$$\tilde{\mathbf{X}} = \sum_{i=1}^{m} \frac{(\mathbf{b}, \mathbf{e}_i) + (\boldsymbol{\eta}, \mathbf{e}_i)}{\lambda_i} \mathbf{e}_i.$$

Le coefficient $\frac{(\boldsymbol{\eta}, \mathbf{e}_i)}{\lambda_i}$ peut s'avérer très grand lorsque λ_i sont petits, ce qui altère fortement le résultat. Parfois, on sait d'avance que dans le développement $\sum_{i=1}^{m} c_i \mathbf{e}_i$ de la solution cherchée, les coefficients c_i correspondant aux λ_i de module petit sont insignifiants. Des mesures s'imposent alors afin de « filtrer » ces composantes de la solution.

Lorsque m sont peu importants, le mode opératoire peut être le suivant : on se donne un $q > 0$, on cherche tous les λ_i et $\mathbf{e}_i$, $i \leqslant q$, et on pose

$$\mathbf{X} \approx \sum_{i=1}^{q} \frac{(\tilde{\mathbf{b}}, \mathbf{e}_i)}{\lambda_i} \mathbf{e}_i.$$

Le choix de q est fonction de l'information supplémentaire sur le problème.

Voici encore deux procédés que nous allons illustrer sur l'exemple de matrices où tous les $\lambda_i > 0$. Dans le premier procédé on se donne un α et on cherche la solution $\mathbf{x}^\alpha$ du système

$$(\alpha E + A)\, \mathbf{x}^\alpha = \tilde{\mathbf{b}};$$

cette solution s'écrit évidemment

$$\mathbf{x}^{\alpha}=\sum_{i=1}^{m}\frac{(\tilde{\mathbf{b}},\ \mathbf{e}_i)}{\lambda_i+\alpha}\mathbf{e}_i.$$

Comme

$$\frac{1}{\lambda_i}-\frac{1}{\lambda_i+\alpha}=\frac{\alpha}{\lambda_i(\lambda_i+\alpha)},$$

la présence du petit paramètre α ne modifie guère les termes associés aux λ_i grands. Par ailleurs, si $\lambda_i \ll \alpha$, on a

$$\left|\frac{(\tilde{\mathbf{b}},\ \mathbf{e}_i)}{\lambda_i+\alpha}\right| \ll \left|\frac{(\tilde{\mathbf{b}},\ \mathbf{e}_i)}{\lambda_i}\right|,$$

ce qui signifie que le fait d'introduire le paramètre α diminue notablement l'importance des termes associés aux λ_i faibles. On choisit la valeur optimale de α en comparant les résultats obtenus pour divers α.

Passons à la deuxième méthode. Résolvons notre système d'équations par un procédé itératif. Considérons les itérations par la formule

$$\mathbf{x}^{n+1}=\mathbf{x}^{n}-\alpha(A\mathbf{x}^{n}-\tilde{\mathbf{b}}) \tag{7}$$

pour, au départ, une approximation $\mathbf{x}^0$. Soit

$$\tilde{\mathbf{b}}=\sum_{i=1}^{m}\beta_i\mathbf{e}_i, \quad \mathbf{x}^{n}=\sum_{i=1}^{m}z_j^{n}\mathbf{e}_j;$$

portons ces expressions dans (7) et égalons les coefficients de $\mathbf{e}_i$; il vient

$$z_i^{n+1}=\alpha\beta_i+(1-\alpha\lambda_i)\,z_i^{n}.$$

En exprimant chaque z_i^k par les précédents, on aboutit à

$$z_i^{n}=\alpha\beta_i+(1-\alpha\lambda_i)\,z_i^{n-1}=\alpha\beta_i+(1-\alpha\lambda_i)\,(\alpha\beta_i+(1-\alpha\lambda_i)\,z_i^{n-2})=\ldots$$

$$\ldots=\alpha\beta_i\sum_{k=0}^{n-1}(1-\alpha\lambda_i)^k+(1-\alpha\lambda_i)^n z_i^0.$$

Si $|1-\alpha\lambda_i|<1$, alors, pour n tendant vers l'infini,

$$\sum_{k=0}^{n-1}(1-\alpha\lambda_i)^k \to \sum_{k=0}^{\infty}(1-\alpha\lambda_i)^k=\frac{1}{\alpha\lambda_i}, \quad (1-\alpha\lambda_i)^n \to 0,$$

et, partant, $z_i^{n} \to \frac{\beta_i}{\lambda_i}$. Soit $\alpha \approx (\max \lambda_i)^{-1}$, i. e. relativement petit. Lorsque λ_i sont importants, la quantité $(1-\alpha\lambda_i)^n$ tend rapidement vers zéro avec la croissance de n et z_i^{n} est voisin de sa valeur limite

$\frac{\beta_i}{\lambda_i}$. D'autre part, on réussit parfois à choisir une approximation initiale telle que z_i^0 soient relativement petits pour λ_i faibles. Alors, lorsque n est peu important, les coefficients de z_i^n correspondant à ces λ_i ne sont pas encore élevés au point d'être inacceptables et l'approximation obtenue peut s'avérer admissible.

Dans d'autres cas le problème se résout par minimisation d'une fonctionnelle proche de $F(\mathbf{x}) = (A\mathbf{x}, \mathbf{x}) - 2(\mathbf{b}, \mathbf{x})$, par exemple $F(\mathbf{x}) + \alpha(\mathbf{x}, \mathbf{x})$ avec $\alpha > 0$ petit.

L'efficacité des techniques décrites dépend essentiellement, lorsqu'elles s'appliquent à des matrices A non symétriques, de la structure de la forme de Jordan et de plusieurs autres propriétés des matrices. La solution s'obtient souvent en l'occurrence en minimisant la fonctionnelle

$$(A\mathbf{x} - \mathbf{b}, A\mathbf{x} - \mathbf{b}) + \alpha(\mathbf{x}, \mathbf{x})$$

pour $\alpha > 0$ petits ; la valeur nécessaire de α est toujours déterminée par la comparaison des calculs avec des α différents.

Un autre groupe de méthodes est basé sur la représentation de la matrice A du système par

$$A = G\Lambda P,$$

G et P étant orthogonales et Λ une matrice bidiagonale dont les éléments λ_{ij} peuvent être non nuls lorsque $j = i$ et $j = i + 1$.

La plupart des techniques décrites permettant de résoudre des systèmes d'équations de matrice mal conditionnée se rapportent aux *méthodes de régularisation.*

PROBLÈME. Soit $A^{(m)} = [a_{ij}^m]$ une $m \times m$-matrice d'éléments

$$a_{ij}^m = \begin{cases} p & \text{pour } j = i, \\ q & \text{pour } j = i+1, \\ 0 & \text{pour } j \neq i,\ i+1. \end{cases}$$

1) Calculer $(A^{(m)})^{-1}$ et démontrer que, pour $|q| < |p|$, les matrices $A^{(m)}$ sont bien conditionnées et que leur conditionnement est mauvais pour $|q| > |p|$ et m grands.

2) Expliciter en fonction du deuxième membre la solution du système $A^{(m)}\mathbf{x} = \mathbf{b}$.

3) Expliciter en $\mathbf{b}$ le vecteur $\mathbf{x}_\alpha$ qui minimise la fonctionnelle

$$(A^{(m)}\mathbf{x} - \mathbf{b}, A^{(m)}\mathbf{x} - \mathbf{b}) + \alpha(\mathbf{x}, \mathbf{x}).$$

4) Exprimer par une relation l'effet d'une telle régularisation.

Il y a une autre raison d'éviter la symétrisation de matrices dont on a parlé au § 7. Voyons ce que donne une opération de symétrisation formelle sur une matrice symétrique. On a alors $A = A^*$ et $A^*A = A^2$. Si une matrice est élevée au carré, il en est de même

de ses valeurs propres; on a donc pour $\| A \| = \| A \|_3$:

$$\nu(A^2) = \frac{\max |\lambda_{A^2}|}{\min |\lambda_{A^2}|} = \frac{(\max |\lambda_A|)^2}{(\min |\lambda_A|)^2} = (\nu(A))^2.$$

Si $\nu(A) = 1$, alors $\max |\lambda_A| = \min |\lambda_A|$ et toutes les valeurs propres de A ont un même module, ce qui donne $A = \text{const}\, E$. Sauf ce cas particulier, $\nu(A) > 1$ et, par conséquent, $\nu(A^2) > \nu(A)$. Ainsi, la symétrisation d'une matrice symétrique augmente le nombre de conditionnement. Cette dernière grandeur étant une fonction continue de la matrice, la même opération sur une matrice proche de la matrice symétrique conduit également à une augmentation du nombre de conditionnement. Il existe donc une classe non vide de matrices non symétriques pour lesquelles l'algorithme de symétrisation augmente le nombre de conditionnement.

PROBLÈME. Existe-t-il des matrices non symétriques telles que $\nu(A^2) < (\nu(A))^2$?

§ 14. Problème de valeurs propres

L'information nécessaire sur les valeurs et vecteurs propres varie selon le cas, d'où la diversité de problèmes et de modes opératoires.

1. La résolution de certains problèmes de mécanique, physique, chimie exige la connaissance de toutes les valeurs propres (et, parfois, de tous les vecteurs propres) de certaines matrices. C'est le *problème global de valeurs propres*. Aujourd'hui, les programmes standard le résolvent pour des matrices de dimensions moyennes (quelques centaines).

2. On demande souvent, en physique nucléaire en particulier, de trouver la valeur propre de plus grand ou de plus petit module. Cela équivaut à rechercher les valeurs propres de matrices de dimension en 10^3-10^6 et plus. Pour des matrices de dimension moindre on utilise surtout des méthodes itératives et pour celles dont la dimension est encore plus importante des procédés probabilistes.

3. L'étude de processus oscillatoires implique parfois la connaissance de deux valeurs propres de plus grand module, la plus petite de ces deux valeurs pouvant être trouvée avec une précision moins bonne.

4. Toujours dans ce domaine, on demande de rechercher une valeur propre qui soit le plus proche d'un nombre donné λ^0 ou bien on demande de déterminer la distance séparant ce nombre du spectre de la matrice.

En raisonnant formellement, on dirait que tous les problèmes restants, qu'on appelle *problèmes locaux de valeurs propres*, sont un cas particulier du problème global, si bien qu'on pourrait se borner

à discuter les méthodes de résolution de ce dernier. Cette approche augmente cependant plus que de raison le volume de calculs. En discutant la position de problèmes concrets relatifs à l'obtention des valeurs propres des matrices, on consacre fréquemment des efforts considérables à établir le plus petit volume d'information nécessaire sur le spectre de la matrice.

La résolution des problèmes 2-4 se ramène d'habitude à la recherche de la valeur propre de plus grand module d'une matrice $B = g(A)$ telle que cette valeur propre corresponde à la valeur propre cherchée de la matrice A.

Considérons le cas où la matrice a toutes ses valeurs propres réelles. Afin de trouver la plus grande ou la plus petite valeur propre de A on prend $g(A) = A + cE$; il est évident que pour c positifs (resp. négatifs) suffisamment élevés il correspond à la valeur propre de plus grand module de $A + cE$ la plus grande (resp. plus petite) valeur propre de A. Dans le cas 4, il existe un c tel que la valeur propre de plus grand module de la matrice $E - c(A - \lambda^0 E)^2$ corresponde à la valeur propre cherchée de A. Parfois on peut prendre pour matrice $g(A)$ la matrice $(A - \lambda^0 E)^{-1}$. Elle ne s'explicite alors pas et les vecteurs $(A - \lambda^0 E)^{-1}\mathbf{y}$ qui sont indispensables aux calculs sont trouvés en résolvant le système $(A - \lambda^0 E)\,\mathbf{x} = \mathbf{y}$.

Considérons un problème type de recherche de deux valeurs propres de plus grand module de la matrice A. Nous simplifierons l'exposé en supposant l'existence d'un système complet de vecteurs propres $\mathbf{e}_i$: $A\mathbf{e}_i = \lambda_i \mathbf{e}_i$, $\|\mathbf{e}_i\| = 1$; soit $|\lambda_1| > |\lambda_2| \geqslant \ldots \geqslant |\lambda_m|$. Donnons-nous un certain vecteur $\mathbf{x}^0$ et calculons de proche en proche les vecteurs $\mathbf{x}^{n+1} = A\mathbf{x}^n$. Mettons $\mathbf{x}^0$ sous la forme $\mathbf{x}^0 = \sum_{i=1}^{m} c_i \mathbf{e}_i$; on a l'égalité évidente

$$\mathbf{x}^n = \sum_{i=1}^{m} c_i A^n \mathbf{e}_i = \sum_{i=1}^{m} c_i \lambda_i^n \mathbf{e}_i.$$

Il en résulte les relations

$$\mathbf{x}^n = c_1 \lambda_1^n \mathbf{e}_1 + O(|\lambda_2|^n),$$

$$(\mathbf{x}^n, \mathbf{x}^n) = (c_1 \lambda_1^n \mathbf{e}_1 + O(|\lambda_2|^n),\ c_1 \lambda_1^n \mathbf{e}_1 + O(|\lambda_2|^n)) =$$

$$= |c_1|^2 |\lambda_1|^{2n} + O(|\lambda_1|^n |\lambda_2|^n),$$

$$(\mathbf{x}^{n+1}, \mathbf{x}^n) = (c_1 \lambda_1^{n+1} \mathbf{e}_1 + O(|\lambda_2|^{n+1}),\ c_1 \lambda_1^n \mathbf{e}_1 + O(|\lambda_2|^n)) =$$

$$= \lambda_1 |c_1|^2 |\lambda_1|^{2n} + O(|\lambda_1|^n |\lambda_2|^n).$$

Posons

$$\lambda_1^{(n)} = \frac{(\mathbf{x}^{n+1}, \mathbf{x}^n)}{(\mathbf{x}^n, \mathbf{x}^n)};$$

les dernières relations nous donnent alors si $c_1 \neq 0$:

$$\lambda_1^{(n)} = \frac{\lambda_1 |c_1|^2 |\lambda_1|^{2n} + O(|\lambda_1|^n |\lambda_2|^n)}{|c_1|^2 |\lambda_1|^{2n} + O(|\lambda_1|^n |\lambda_2|^n)} =$$

$$= \frac{\lambda_1 \left(1 + O\left(\frac{1}{|c_1|^2} \left|\frac{\lambda_2}{\lambda_1}\right|^n\right)\right)}{1 + O\left(\frac{1}{|c_1|^2} \left|\frac{\lambda_2}{\lambda_1}\right|^n\right)} = \lambda_1 + O\left(\left|\frac{\lambda_2}{\lambda_1}\right|^n\right). \quad (1)$$

PROBLÈME. Démontrer que dans le cas d'une matrice symétrique

$$\lambda_1^{(n)} = \lambda_1 + O\left(\left|\frac{\lambda_2}{\lambda_1}\right|^{2n}\right).$$

On a, en plus de (1),

$$\|x^n\| = \sqrt{(x^n, x^n)} = |c_1| |\lambda_1|^n + O(|\lambda_2|^n),$$

$$e_1^{(n)} = \frac{x^n}{\|x^n\|} = \frac{\sum_{i=1}^{m} c_i \lambda_i^n e_i}{|c_1 \lambda_1^n| + O(|\lambda_2|^n)} =$$

$$= \frac{\frac{c_1 \lambda_1^n}{|c_1 \lambda_1^n|} e_1 + O\left(\left|\frac{\lambda_2}{\lambda_1}\right|^n\right)}{1 + O\left(\left|\frac{\lambda_2}{\lambda_1}\right|^n\right)} = e^{i\varphi_n} e_1 + O\left(\left|\frac{\lambda_2}{\lambda_1}\right|^n\right);$$

ici $\varphi_n = \arg(c_1 \lambda_1^n)$.

Donc, ce processus itératif aboutit lui aussi à un vecteur propre associé à λ_1.

Il se peut que la matrice A possède deux valeurs propres de plus grand module, à savoir

$$\lambda_1 \neq \lambda_2, \quad |\lambda_1| = |\lambda_2| > |\lambda_3| \geqslant \ldots$$

La quantité $\lambda_1^{(n)}$ ne se stabilise alors évidemment que dans le cas particulier de c_1 ou $c_2 = 0$. Si l'on sait d'avance que ces valeurs propres sont au nombre de deux, on peut les déterminer, ainsi que les vecteurs propres correspondants, en analysant le comportement de $\lambda_1^{(n)}$ et $e_1^{(n)}$.

Etudions le cas typique où λ_1, λ_2 sont réels, $\lambda_1 > 0$, $\lambda_1 = -\lambda_2$. Alors

$$x^n = c_1 \lambda_1^n e_1 + c_2 (-\lambda_1)^n e_2 + O(|\lambda_3|^n),$$

$$x^{n+2} = c_1 \lambda_1^{n+2} e_1 + c_2 (-\lambda_1)^{n+2} e_2 + O(|\lambda_3|^n) =$$

$$= \lambda_1^2 (c_1 \lambda_1^n e_1 + c_2 (-\lambda_1)^n e_2) + O(|\lambda_3|^n).$$

Il en résulte

$$\tilde{\lambda}^{(n)} = \frac{(x^{n+2}, x^n)}{(x^n, x^n)} = \lambda_1^2 + O\left(\left|\frac{\lambda_3}{\lambda_1}\right|^n\right)$$

et les valeurs approchées $\lambda_1^{(n)} > 0$ et $\lambda_2^{(n)} < 0$. On a

$$z_1^{n+1} = x^{n+1} + \lambda_1^{(n)} x^n = 2c_1\lambda_1^{n+1} e_1 + O(|\lambda_3|^n),$$

et donc

$$e_1^{(n)} = \frac{z_1^{n+1}}{\| z_1^{n+1} \|} = e^{i\psi_n} e_1 + O\left(\left|\frac{\lambda_3}{\lambda_1}\right|^n\right).$$

De même

$$z_2^{n+1} = x^{n+1} + \lambda_2^{(n)} x^n = 2c_2\lambda_1^{n+1} e_2 + O(|\lambda_3|^n)$$

et

$$e_2^{(n)} = \frac{z_2^{n+1}}{\| z_2^{n+1} \|} = e^{i\omega_n} e_2 + O\left(\left|\frac{\lambda_3}{\lambda_1}\right|^n\right).$$

Si λ_i sont les valeurs propres de A, la conjuguée A^* a $\overline{\lambda}_i$ pour valeur propre; alors, si $Ae_i = \lambda_i e_i$, $A^* g_j = \overline{\lambda}_j g_j$, $\lambda_i \neq \lambda_j$, on a $(e_i, g_j) = 0$.

Aussi, ayant déterminé la valeur approchée $\widetilde{l}_1$ du vecteur l_1 pour $|\lambda_1| > |\lambda_2| > |\lambda_3| \geqslant \ldots$, la valeur propre λ_2 s'obtient de la façon suivante. Un processus itératif analogue nous fournit l'approximation $\widetilde{g}_1$ du vecteur g_1, $(\widetilde{e}_1, \widetilde{g}_1) = 1$. Itérons ensuite par la formule $y^{n+1} = Ay^n$; afin d'éliminer la composante proportionnelle à e_1, rendons de temps en temps les y^n orthogonaux au vecteur $\widetilde{g}_1$, c'est-à-dire que pour vecteur de départ nous prenons dans les itérations suivantes le vecteur

$$\widetilde{y}_n = y^n - (y^n, \widetilde{g}_1)\widetilde{e}_1$$

au lieu de y^n. Il est naturel que la convergence du processus soit meilleure si le terme d_2e_2 de l'approximation initiale $y_0 = \sum_{i=1}^{m} d_i e_i$ est prépondérant. Prenons une certaine approximation x^l du processus décrit de recherche de λ_1 et e_1 et rendons-la orthogonale à g_1. Acceptons comme approximation initiale le vecteur $y^0 = x^l - (x^l, \widetilde{g}_1)\widetilde{e}_1$ ainsi obtenu. Gardons-nous de prendre l petit (les composantes $c_j\lambda_j^l e_j$, $j > 2$, ne seront pas encore petites devant $c_2\lambda_2^l e_2$) ou très grand (la composante $c_2\lambda_2^l e_2$ sera alors faible devant l'erreur de calcul).

Selon certains auteurs, si $c_1 = 0$, le processus itératif décrit ne devrait pas donner d'approximation convergeant vers la valeur propre de plus grand module. Cependant, les arrondis font apparaître une composante proportionnelle à e_1, et le résultat cherché est forcément obtenu. Ceci arrive effectivement en calcul manuel avec une précision peu élevée. Par contre, plusieurs itérations exécutées sur une machine moderne très capacitive peuvent donner lieu à une erreur de calcul dont la répercussion sera insignifiante, tandis que $\sum_{i=3}^{m} c_i\lambda_i^n e_i$

sera négligeable devant $c_2\lambda_2^n \mathbf{e}_2$. Alors $A\mathbf{x}^n \approx$ const $\mathbf{x}^n$ et on pourrait penser à tort avoir trouvé la première valeur propre cherchée. Dans plusieurs cas, l'apparition d'une composante proportionnelle à $\mathbf{e}_1$ n'est pas forcément obligatoire même en présence d'arrondis. Dans le problème de valeurs propres pour les opérateurs différentiels et intégraux on voit parfois apparaître des matrices A douées de propriétés spéciales; ainsi, on a fréquemment

$$a_{ij} = a_{m+1-i,\, m+1-j}, \tag{2}$$

quels que soient i, j. Pour fixer les idées, supposons m pair. Nous dirons que les vecteurs $\mathbf{x}$ sont pairs si leurs composantes sont reliées par $x_i = x_{m+1-i}$, $i = 1, \ldots, m$, et impairs si $x_i = -x_{m+1-i}$. Dans la condition (2) le vecteur $A\mathbf{x}$ est pair ou impair avec $\mathbf{x}$. C'est pourquoi les sous-espaces de vecteurs pairs ou impairs sont des sous-espaces propres de l'opérateur A. Il existe donc un système complet de vecteurs propres appartenant à ces sous-espaces, c'est-à-dire qui sont ou pairs ou impairs. Cette circonstance est essentielle: si un vecteur $\mathbf{x}^0$ est pair (resp. impair), il en est de même de tous les $\mathbf{x}^n$ si bien qu'en cherchant chaque $\mathbf{x}^n$ suivant il suffit de déterminer la première moitié de ses composantes. On réunit de plus les coefficients correspondant aux composantes x_i et x_{m+1-i}. Pour $\mathbf{x}^0$ pair, les composantes de $\mathbf{x}^n$ se calculent par exemple à l'aide des formules

$$x_i^{n+1} = \sum_{j=1}^{m/2} (a_{ij} + a_{i,\, m+1-j})\, x_i^n.$$

Avec ces itérations nous restons dans les limites des sous-espaces de vecteurs pairs ou impairs respectivement. Aussi, pour $\mathbf{x}^0$ et $\mathbf{e}_1$ éléments de sous-espaces différents, la composante proportionnelle à $\mathbf{e}_1$ n'apparaît pas. A part qu'elle permet de diminuer directement le volume de calculs qu'exige la recherche de chaque $\mathbf{x}^n$, la propriété (2) rend un autre service. Un cas assez répandu est celui où les vecteurs propres de numéros impairs petits sont pairs et ceux de numéros pairs petits impairs; on admet que $|\lambda_1| > |\lambda_2| > |\lambda_3| > \ldots$ Alors, si $\mathbf{x}^0$ est pair, on a toujours

$$\mathbf{x}^n = c_1\lambda_1^n \mathbf{e}_1 + c_3\lambda_3^n \mathbf{e}_3 + \ldots = c_1\lambda_1^n \mathbf{e}_1 + O(|\lambda_3|^n),$$

par conséquent,

$$\lambda_1^{(n)} - \lambda_1 = O\left(\left|\frac{\lambda_3}{\lambda_1}\right|^n\right) \quad \text{avec} \quad \lambda_1^{(n)} \frac{(\mathbf{x}^{n+1}, \mathbf{x}^n)}{(\mathbf{x}^n, \mathbf{x}^n)}.$$

De même, pour $\mathbf{x}^0$ impair,

$$\mathbf{x}^n = c_2\lambda_2^n \mathbf{e}_2 + O(|\lambda_4|^n)$$

et

$$\lambda_2^{(n)} - \lambda_2 = O\left(\left|\frac{\lambda_4}{\lambda_2}\right|^n\right) \quad \text{avec} \quad \lambda_2^{(n)} = \frac{(\mathbf{x}^{n+1}, \mathbf{x}^n)}{(\mathbf{x}^n, \mathbf{x}^n)}.$$

Ceci étant, on n'a pas besoin de réprimer la composante proportionnelle à $\mathbf{e}_1$.

Si $|\lambda_1| > 1$, alors $\|\mathbf{x}^n\|$ tend vers l'infini avec n et, pour n suffisamment grand, il y a dépassement de capacité des nombres et arrêt de la machine. Si, par contre, $|\lambda_1| < 1$, alors $\|\mathbf{x}^n\| \to 0$, et, l'exposant étant fini, il se peut que, à partir d'un certain n, $\mathbf{x}^n \equiv 0$. On l'évite en multipliant de temps en temps le vecteur $\mathbf{x}^n$ par un facteur convenablement choisi.

On évalue l'erreur et on accélère la convergence des processus itératifs par le procédé δ^2 et par d'autres méthodes analogues aux méthodes d'accélération de la convergence pour les systèmes d'équations linéaires. Ce peut être par exemple des itérations de la forme $x^{n+1} = g_k(A)x^n$ où le polynôme $g_k(A)$ est choisi en fonction de l'information dont on dispose sur le spectre de la matrice A.

Puisque

$$\max \lambda_A^i = \sup_x \frac{(Ax, x)}{(x, x)}, \quad \min \lambda_A^i = \inf_x \frac{(Ax, x)}{(x, x)}$$

pour $A = A^*$, certaines techniques de recherche de la plus grande et de la plus petite valeur propre de A s'appuient sur la recherche des points stationnaires de la fonctionnelle

$$\Phi(x) = \frac{(Ax, x)}{(x, x)}.$$

PROBLÈME. Soit $\lambda_1 \approx 5$, $1 \leqslant \lambda_i \leqslant 3$ pour $i = 2, \ldots, m$. Construire un processus itératif de la forme $x^{n+1} = (A + cE)x^n$ afin d'obtenir λ_1 avec la meilleure vitesse de convergence pour l'information donnée.

Résoudre le même problème si $\lambda_1 \approx 1$, $2 \leqslant |\lambda_i| \leqslant 3$, $i = 2, \ldots, m$.

§ 15. Résolution du problème global de valeurs propres d'une matrice symétrique par la méthode des rotations

Toute matrice symétrique A peut s'écrire

$$A = U^* \Lambda U, \tag{1}$$

où U est orthogonale et Λ diagonale et telle que les éléments diagonaux $\lambda_1, \ldots, \lambda_m$ sont valeurs propres de A. (1) entraîne

$$AU^* = U^* \Lambda$$

En développant cette égalité suivant les colonnes, on constate que chaque i-ième colonne de la matrice U^* est un vecteur propre associé à la valeur propre λ_i.

La transformation de (1) donne

$$UAU^* = \Lambda.$$

Formons une suite de matrices orthogonales $V_1 \ldots V_n$ telle que si $W_n = V_n \ldots V_1$, $n \to \infty$, on ait

$$W_n A W_n^* \to \Lambda.$$

Notons $V_{kl}(\varphi)$ une matrice d'éléments $v_{kk} = v_{ll} = \cos \varphi$, $v_{kl} = -v_{lk} = \sin \varphi$, $v_{ii} = 1$ avec $i \neq k$, $i \neq l$, $v_{ij} = 0$ pour les (i, j) restants. Il est évident qu'elle est orthogonale. Soit B une

matrice symétrique. Considérons la matrice

$$B(\varphi) = V_{kl}(\varphi)\, B V_{kl}^{*}(\varphi). \tag{2}$$

LEMME. *Pour toute matrice B symétrique on peut indiquer une matrice $V_{kl}(\varphi_0)$ telle que*

$$\sum_{i \neq j} |b_{ij}^2(\varphi_0)|^2 \leqslant \left(1 - \frac{2}{m(m-1)}\right) \sum_{i \neq j} |b_{ij}|^2.$$

DÉMONSTRATION. L'égalité

$$B(0) = B$$

est évidente.

Les matrices $B(0)$ et $B(\varphi)$ sont semblables et leurs polynômes caractéristiques

$$\lambda^m - s_1(\varphi)\lambda^{m-1} + s_2(\varphi)\lambda^{m-2} - \ldots = 0$$

ne dépendent par conséquent pas de φ.

Ici

$$s_1(\varphi) = \sum_{i=1}^{m} b_{ii}(\varphi), \quad s_2(\varphi) = \sum_{i<j} (b_{ii}(\varphi)\, b_{jj}(\varphi) - (b_{ij}(\varphi))^2);$$

en explicitant $s_2(\varphi)$ nous avons utilisé la symétrie de $B(\varphi)$. Il en découle que

$$s(\varphi) = s_1^2(\varphi) - 2s_2(\varphi) = \sum_{i,j} (b_{ij}(\varphi))^2$$

ne dépend pas de φ :

$$s(\varphi) = s(0). \tag{3}$$

Soit $|b_{kl}| = \max\limits_{i \neq j} |b_{ij}|$, $k < l$.

Introduisons la notation

$$B_{kl}(\varphi) = \begin{pmatrix} b_{kk}(\varphi) & b_{kl}(\varphi) \\ b_{lk}(\varphi) & b_{ll}(\varphi) \end{pmatrix}.$$

Une multiplication directe des matrices à droite dans (2) et la comparaison des éléments appartenant aux k-ième et l-ième lignes et aux k-ième et l-ième colonnes fournissent les relations

$$B_{kl}(\varphi) = G(\varphi)\, B_{kl}(0)\, G^{*}(\varphi) \tag{4}$$

avec

$$G(\varphi) = \begin{pmatrix} \cos\varphi & \sin\varphi \\ -\sin\varphi & \cos\varphi \end{pmatrix}.$$

L'égalité (3) est également satisfaite par les matrices $B_{kl}(\varphi)$, $B_{kl}(0)$; c'est pourquoi

$$(b_{kk}(\varphi))^2 + 2(b_{kl}(\varphi))^2 + (b_{ll}(\varphi))^2 = (b_{kk})^2 + 2(b_{kl})^2 + (b_{ll})^2. \tag{5}$$

En additionnant (5) et

$$(b_{qq}(\varphi))^2 = (b_{qq})^2, \quad q \neq k, l,$$

il vient

$$\sum_{q=1}^{m} (b_{qq}(\varphi))^2 + 2(b_{kl}(\varphi))^2 = \sum_{q=1}^{m} (b_{qq})^2 + 2(b_{kl})^2.$$

Retranchons cette dernière égalité de (3) :

$$\sum_{i \neq j} (b_{ij}(\varphi))^2 - 2(b_{kl}(\varphi))^2 = \sum_{i \neq j} (b_{ij})^2 - 2(b_{kl})^2. \tag{6}$$

Les relations (4) entraînent immédiatement

$$b_{kl}(\varphi) = b_{kl} \cos 2\varphi - \frac{1}{2}(b_{kk} - b_{ll}) \sin 2\varphi.$$

Soit

$$\varphi_0 = \frac{1}{2} \operatorname{arctg} \frac{2b_{kl}}{b_{kk} - b_{ll}};$$

alors $b_{kl}(\varphi_0) = 0$ et (6) devient

$$\sum_{i \neq j} (b_{ij}(\varphi_0))^2 = \sum_{i \neq j} (b_{ij})^2 - 2(b_{kl})^2.$$

La somme $\sum_{i \neq j} b_{ij}^2$ se compose de $m(m-1)$ termes b_{ij}^2; comme $(b_{kl})^2 = \max_{i \neq j} (b_{ij})^2$, on a

$$2(b_{kl})^2 \geqslant \frac{2}{m(m-1)} \sum_{i \neq j} (b_{ij})^2.$$

D'où l'estimation demandée

$$\sum_{i \neq j} (b_{ij}(\varphi_0))^2 \leqslant \left(1 - \frac{2}{m(m-1)}\right) \sum_{i \neq j} (b_{ij})^2.$$

Le lemme est démontré.

Formons une suite de matrices V_n et $A^{(n)} = [a_{ij}^{(n)}]$ d'après la règle

$$A^{(1)} = V_1 A V_1^*, \ \ldots, \ A^{(n)} = V_n A^{(n-1)} V_n^*.$$

La matrice $V_n = B_{kl}(\varphi_0)$ est construite, conformément au lemme, en fonction de $A^{(n-1)}$ de telle façon que

$$\sum_{i \neq j} (a_{ij}^{(n)})^2 \leqslant \left(1 - \frac{2}{m(m-1)}\right) \sum_{i \neq j} (a_{ij}^{(n-1)})^2. \tag{7}$$

En exprimant chaque matrice $A^{(p)}$ par la précédente, il vient

$$A^{(n)} = V_n \ldots V_1 A V_1^* \ldots V_n^*,$$

ou bien $A^{(n)} = W_n A W_n^*$ avec $W_n = V_n \ldots V_1$. Conformément à l'estimation (7), on a

$$\sum_{i \neq j} (a_{ij}^{(n)})^2 \leqslant \left(1 - \frac{2}{m(m-1)}\right)^n \sum_{i \neq j} (a_{ij})^2.$$

Soit $\Lambda^{(n)}$ une matrice obtenue de $A^{(n)}$ en remplaçant tous les éléments extradiagonaux par zéros. Comme les matrices $\Lambda^{(n)}$ et $A^{(n)}$ sont voisines pour n grands, il en est de même des coefficients de leurs polynômes caractéristiques. On sait que les racines d'un polynôme sont fonction continue des coefficients dans le domaine où le coefficient dominant est différent de zéro ; le coefficient dominant du polynôme caractéristique vaut 1 ; aussi les valeurs propres de $\Lambda^{(n)}$ et $A^{(n)}$ sont-elles voisines. Les valeurs propres de $\Lambda^{(n)}$ sont éléments diagonaux de $A^{(n)}$ et celles des matrices semblables $A^{(n)}$ et A coïncident. De sorte que les quantités $a_{(ii)}^{(n)}$ approchent les valeurs propres de la matrice A.

Donnons une autre démonstration en appréciant cette fois l'erreur commise. L'application à deux reprises de l'inégalité

$$\left|\sum_i \alpha_i\beta_i\right| \leqslant \sqrt{\left(\sum_i |\alpha_i|^2\right)\left(\sum_i |\beta_i|^2\right)}$$

fournit

$$|(D\mathbf{x}, \mathbf{x})| = \left|\sum_i \left(\sum_j d_{ij}x_j\right) x_i\right| \leqslant \sum_i \left|\sum_j d_{ij}x_j\right| \cdot |x_i| \leqslant$$

$$\leqslant \sum_i \sqrt{\left(\sum_j |d_{ij}|^2\right)\left(\sum_j |x_j|^2\right)}\,|x_i| = \sqrt{\sum_j |x_j|^2}\left(\sum_i \sqrt{\sum_j |d_{ij}|^2}\,|x_i|\right) \leqslant$$

$$\leqslant \sqrt{\sum_j |x_j|^2}\,\sqrt{\left(\sum_i \left(\sum_j |d_{ij}|^2\right)\right)\sum_i |x_i|^2} = \sqrt{\sum_{ij} |d_{ij}|^2}\,(\mathbf{x}, \mathbf{x}) \qquad (8)$$

pour n'importe quelle matrice D.

Rappelons-nous la formulation de la propriété minimaximale des valeurs propres. Supposons que A est symétrique et $\lambda_1, \ldots, \lambda_m$ sont ses valeurs propres ordonnées dans l'ordre de décroissance. Alors

$$\lambda_k = \inf_{y_1, \ldots, y_{k-1}} \sup_{(\mathbf{x}, y_1)=\ldots=(\mathbf{x}, y_{k-1})=0,\ \mathbf{x}\neq 0} \frac{(A\mathbf{x}, \mathbf{x})}{(\mathbf{x}, \mathbf{x})}.$$

Quel que soit $\mathbf{x}$, on a selon (8)

$$|(A^{(n)}\mathbf{x}, \mathbf{x}) - (\Lambda^{(n)}\mathbf{x}, \mathbf{x})| = |((A^{(n)} - \Lambda^{(n)})\,\mathbf{x}, \mathbf{x})| \leqslant \sqrt{\sum_{i\neq j} |a_{ij}^{(n)}|^2}\,(\mathbf{x}, \mathbf{x}). \qquad (9)$$

Soient λ_k les valeurs propres de A et, partant, de $A^{(n)}$ numérotées dans l'ordre décroissant et $\lambda_k^{(n)}$ celles de la matrice $\Lambda^{(n)}$, i.e. ses éléments diagonaux, numérotés de façon identique. Conformément à (9), on a

$$\left|\frac{(A^{(n)}\mathbf{x}, \mathbf{x})}{(\mathbf{x}, \mathbf{x})} - \frac{(\Lambda^{(n)}\mathbf{x}, \mathbf{x})}{(\mathbf{x}, \mathbf{x})}\right| \leqslant \sigma_n = \sqrt{\sum_{i\neq j} |a_{ij}^{(n)}|^2}.$$

Si deux fonctions diffèrent de ε au plus, la différence de leurs bornes supérieures (resp. inférieures) prises sur un certain ensemble vaut également ε au plus. D'où les inégalités successives

$$\left|\sup_{(\mathbf{x}, y_1)=\ldots=(\mathbf{x}, y_{k-1})=0,\ \mathbf{x}\neq 0} \frac{(A^{(n)}\mathbf{x}, \mathbf{x})}{(\mathbf{x}, \mathbf{x})} - \sup_{(\mathbf{x}, y_1)=\ldots=(\mathbf{x}, y_{k-1})=0,\ \mathbf{x}\neq 0} \frac{(\Lambda^{(n)}\mathbf{x}, \mathbf{x})}{(\mathbf{x}, \mathbf{x})}\right| \leqslant$$

$\leqslant \sigma_n$ quels que soient $y, \ldots, y_{k-1}$

et

$$|\lambda_k^{(n)} - \lambda_k| \leqslant \sigma_n \quad \text{quel que soit } k.$$

Les colonnes de la matrice W_n sont les approximations des vecteurs propres de la matrice A.

Toutes les colonnes du produit des matrices $V_{kl}(\varphi_0)$ et $A^{(n)}$, à l'exception de la k-ième et de la l-ième, coïncident avec les colonnes correspondantes de $A^{(n)}$; les k-ième et l-ième colonnes de la nouvelle matrice sont combinaisons linéaires des mêmes colonnes de la matrice A. La multiplication ultérieure par $V_{kl}^*(\varphi)$ qui change la k-ième et la l-ième ligne exige le même nombre d'opérations. Le nombre total d'opérations arithmétiques nécessaires pour obtenir chaque $A^{(n)}$ suivante est en gros $6m$.

L'algorithme proposé présente l'inconvénient de recommander de choisir à chaque pas l'élément de plus grand module de $A^{(n)}$, ce qui équivaut à $\asymp m^2$ opérations. Or, cette condition n'est pas obligatoire.

Dans certaines mises en œuvre de l'algorithme on charge en mémoire les sommes $s_i = \sum_{j \neq i} |a_{ij}|^2$ et on choisit l'élément de plus grand module dans la ligne contenant la valeur maximale de s_i; l'estimation (7) reste en vigueur. Dans d'autres cas, les transformations analogues correspondant aux divers éléments de la matrice $a_{ij}^{(n)}$, $j > i$, sont effectuées de façon cyclique. Certains algorithmes sont à convergence quadratique. Après un cycle de m $(m - 1)/2$ pas on obtient une matrice $A^{(n+(m(m-1)/2))}$ telle que

$$\sum_{i \neq j} |a_{ij}^{(n+(m(m-1)/2))}|^2 \leqslant c\Big(\sum_{i \neq j} |a_{ij}^{(n)}|^2\Big)^2.$$

Avec une approximation assez bonne, ceci garantit une amélioration rapide du résultat.

Le problème global de valeurs propres d'une matrice non symétrique se résout par des algorithmes plus compliqués, mais d'une structure toujours homogène.

Bibliographie du chapitre VI

I [1, 2, 9, 12], II [5, 17, 46]; §§ 4, 5: I [4], II [6] *); § 7: II [4, 5, 17, 27, 28]; § 10: III [30]; § 11: II [2, 18]; § 12: II [5]; III [26]; § 13: II [15], III [46]; §§ 14, 15: I [2], II [46].

*) Voir également V. Riabenki « A propos de la stabilité des algorithmes itératifs de résolution d'équations aux différences non auto-adjointes », Comptes rendus de l'Académie des Sciences de l'U.R.S.S., 1970, 193, n° 3 (éd. russe).

CHAPITRE VII

RÉSOLUTION DES SYSTÈMES D'ÉQUATIONS NON LINÉAIRES ET DES PROBLÈMES D'OPTIMISATION

La résolution des problèmes d'optimisation, qu'il s'agisse de la production ou de l'économie en général, de constructions ou d'algorithmes numériques, se ramène mathématiquement à la recherche des extréma de certaines fonctionnelles.

Dans les cas les plus typiques, on a à minimiser une fonction de plusieurs variables dans un domaine G donné par de nombreuses contraintes de type inégalité ou égalité : on désire trouver

$$\inf_{x_1, \ldots, x_m} \Phi(x_1, \ldots, x_m)$$

dans les conditions

$$\varphi_i(x_1, \ldots, x_m) \geqslant 0, \quad i = 1, \ldots, l.$$

$$\psi_i(x_1, \ldots, x_m) = 0, \quad i = 1, \ldots, q.$$

Le même problème se pose lorsqu'on utilise les méthodes variationnelles en physique mathématique et dans d'autres branches des mathématiques appliquées. De nombreux types de problèmes indiqués donnent lieu aux systèmes

$$\mathbf{F}(\mathbf{x}) = 0 : f_i(x_1, \ldots, x_m) = 0, \quad i = 1, \ldots, m;$$

ainsi, dans le chapitre IX on considérera de tels systèmes lorsqu'on aura à résoudre des problèmes aux limites.

Il y a parfois intérêt à ramener un problème à un autre.

Si $\Psi(y_1, \ldots, y_m) > 0$ avec $(y_1, \ldots, y_m) \neq (0, \ldots, 0)$, $\Psi(0, \ldots, 0) = 0$, résoudre le système $\mathbf{F}(\mathbf{x}) = 0$ équivaut à minimiser la fonction

$$\Psi(f_1(x_1, \ldots, x_m), \ldots, f_m(x_1, \ldots, x_m)).$$

D'autre part, supposons que $\inf_G \Phi(x_1, \ldots, x_m)$ soit réalisée en un point X intérieur à G et que la fonction Φ soit dérivable en ce point. Alors le point de minimum est solution du système d'équations

$$\Phi'_{x_i} = 0, \quad i = 1, \ldots, m.$$

La possibilité de ramener un problème à tel autre ne veut nullement dire que nous pouvons nous borner à l'un d'eux ; la parenté

de ces problèmes souligne plutôt qu'ils sont également ardus. Car ils sont très difficiles. Vous ne trouverez en effet dans ce chapitre *aucun algorithme universel* de résolution dans le cas d'un m grand. On effectue une « mise au point » théorique et expérimentale des méthodes en vue de les adapter à chaque classe type quant à ses applications.

Si l'on ramène un problème à un autre, c'est pour réduire l'effort qu'exige sa résolution.

Dans le cas des systèmes d'équations non linéaires, on peut par exemple procéder comme suit. On construit une fonctionnelle dont le minimum est atteint sur la solution du système. On se donne ensuite une approximation initiale de ce point et on réalise les itérations par une technique de descente (cf. § 4) jusqu'à obtenir une approximation satisfaisante de la solution. Avec cette approximation au départ, on améliore la précision moyennant une méthode itérative spéciale au problème proposé, par exemple la méthode de Newton (cf. § 2).

Expliquons le pourquoi d'une pareille combinaison de procédés. Appelons *domaine de convergence d'une méthode* l'ensemble de conditions initiales pour lesquelles les itérations de la méthode considérée convergent vers la solution du problème. Nous avons commencé par appliquer les méthodes de descente parce que leur domaine de convergence est en général plus vaste que celui des techniques propres à la résolution de systèmes d'équations. Le fait que ces dernières convergent en général plus rapidement avec une approximation initiale suffisamment bonne explique leur utilisation à l'étape finale du processus itératif.

D'ailleurs, en résolvant des systèmes d'équations linéaires nous avons également constaté que la réduction d'un problème à un autre exige l'élaboration de nouvelles méthodes de résolution du problème initial.

§ 1. Méthode itérative simple et problèmes connexes

Comme dans le cas d'équations linéaires, commençons par la *méthode itérative simple.*

Voici son idée. On met le système sous la forme

$$\mathbf{x} = \mathbf{g}(\mathbf{x}), \tag{1}$$

ou encore

$$x_i = g_i(x_1, \ldots, x_m), \quad i = 1, \ldots, m,$$

et on effectue les itérations à l'aide de la formule

$$\mathbf{x}^{n+1} = \mathbf{g}(\mathbf{x}^n), \tag{2}$$

ou encore

$$x_i^{n+1} = g_i(x_1^n, \ldots, x_m^n), \quad i = 1, \ldots, m.$$

Passons du concret au général et supposons que $\mathscr{X}$ est un espace métrique complet et que $y = g(x)$ est un opérateur de $\mathscr{X}$ dans lui-même.

Considérons le processus itératif

$$x^{n+1} = g(x^n) \tag{3}$$

permettant de résoudre

$$x = g(x). \tag{4}$$

Si pour un certain $q < 1$ l'application $y = g(x)$ vérifie la condition

$$\rho(g(x_1), g(x_2)) \leqslant q\rho(x_1, x_2) \tag{5}$$

quels que soient x_1, x_2, on l'appelle *application contractante*.

THÉORÈME. *Si l'application* $y = g(x)$ *est contractante, l'équation* $x = g(x)$ *possède une solution* X *et*

$$\rho(X, x^n) \leqslant \frac{q^n a}{1-q},$$

$a = \rho(x^1, x^0)$.

DÉMONSTRATION. Selon (5), on a

$$\rho(x^{n+1}, x^n) = \rho(g(x^n), g(x^{n-1})) \leqslant q\rho(x^n, x^{n-1}),$$

aussi $\rho(x^{n+1}, x^n) \leqslant q^n \rho(x^1, x^0) = q^n a$. Si $p > n$, on a la chaîne d'inégalités

$$\rho(x^p, x^n) \leqslant \rho(x^p, x^{p-1}) + \ldots + \rho(x^{n+1}, x^n) \leqslant$$

$$\leqslant q^{p-1}a + \ldots + q^n a \leqslant q^n a \sum_{i=0}^{\infty} q^i = \frac{q^n a}{1-q}. \tag{6}$$

Conformément au critère de Cauchy, la suite x^n tend vers une limite X. Le passage à la limite dans (6), $p \to \infty$, fournit

$$\rho(X, x^n) \leqslant \frac{q^n a}{1-q}.$$

Evidemment

$$\rho(X, g(X)) \leqslant \rho(X, x^{n+1}) + \rho(x^{n+1}, g(X)) =$$

$$= \rho(X, x^{n+1}) + \rho(g(x^n), g(X)) \leqslant \rho(X, x^{n+1}) + q\rho(x^n, X) \leqslant 2\frac{q^{n+1}a}{1-q};$$

n étant quelconque, $\rho(X, g(X)) = 0$, i.e. $X = g(X)$, c.q.f.d.

REMARQUE. Avec $n = 0$, (6) entraîne $\rho(x^p, x^0) \leqslant \frac{a}{1-q}$. Toutes les approximations appartiennent donc au domaine

$$\Omega(x^0, a, q) : \rho(x, x^0) \leqslant \frac{a}{1-q}.$$

Dans la démonstration du théorème $g(x)$ ne s'applique qu'aux éléments de l'ensemble $\Omega(x^0, a, q)$ et la condition de contraction est relative à un seul

couple d'éléments de cet ensemble. Il suffit donc de supposer dans l'énoncé que l'application $g(x)$ est définie sur les éléments de $\Omega(x^0, a, q)$ et satisfait à la condition (5) pour $x_1, x_2 \in \Omega(x^0, a, q)$.

Dans le cas d'une équation scalaire, la méthode itérative simple admet une interprétation géométrique simple. Représentons dans le plan (x, y) la position relative de la courbe $y = g(x)$ et de la

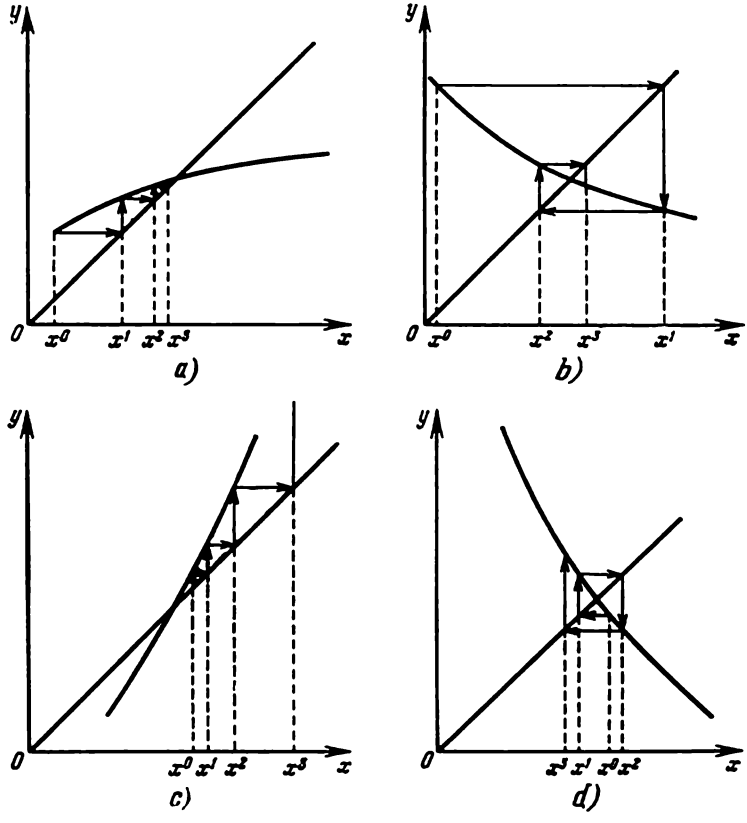


Fig. 7.1.1

droite $y = x$. Leurs points d'intersection correspondent à la solution cherchée. Si par le point $(x^n, x^{n+1}) = (x^n, g(x^n))$ on mène la droite $y = x^{n+1}$ jusqu'à ce qu'elle rencontre la droite $y = x$, puis la droite $x = x^{n+1}$ qui va couper la courbe $y = g(x)$, on obtient le point (x^{n+1}, x^{n+2}). La fig. 7.1.1. représente le comportement des approximations successives si a) $0 < g'(x) < 1$, b) $-1 < g'(x) < 0$, c) $1 < g'(x)$, d) $g'(x) < -1$. Le fait que x^n est monotone pour $g' > 0$ et oscillante lorsque $g' < 0$ est également aisé à établir

à l'aide de

$$x^{n+1} - X = g(x^n) - g(X) \sim g'(X)(x^n - X).$$

Le système d'équations non linéaires $\mathbf{F}(\mathbf{x}) = 0$ se résout par un analogue de la méthode de Seidel dans lequel les composantes d'approximations sont données par les relations

$$\begin{aligned} f_1(x_1^{n+1}, x_2^n, \ldots, x_m^n) &= 0, \\ f_2(x_1^{n+1}, x_2^{n+1}, x_3^n, \ldots, x_m^n) &= 0, \\ \cdots\cdots\cdots\cdots\cdots\cdots & \\ f_m(x_1^{n+1}, \ldots, x_m^{n+1}) &= 0. \end{aligned} \tag{7}$$

Pour trouver chaque x_i^{n+1} il faut en général résoudre l'équation non linéaire

$$f_i(x_1^{n+1}, \ldots, x_{i-1}^{n+1}, x_i^{n+1}, x_{i+1}^n, \ldots, x_m^n) = 0$$

à une inconnue.

Entre les procédés par itérations (2) et (7) se situe une méthode dans laquelle les composantes d'approximations se déterminent par

$$\begin{aligned} x_1^{n+1} &= g_1(x_1^n, \ldots, x_m^n), \\ x_2^{n+1} &= g_2(x_1^{n+1}, x_2^n, \ldots, x_m^n), \\ \cdots\cdots\cdots\cdots\cdots\cdots & \\ x_m^{n+1} &= g_m(x_1^{n+1}, \ldots, x_{m-1}^{n+1}, x_m^n). \end{aligned} \tag{8}$$

Les méthodes (7) et (8) sont d'un usage particulièrement fréquent lors des traitements sur calculateurs analogiques.

Dans un voisinage relativement restreint de la solution X du système on a en approximant par la méthode itérative simple:

$$\mathbf{x}^{n+1} - \mathbf{X} = \mathbf{g}(\mathbf{x}^n) - \mathbf{g}(\mathbf{X}) \approx B(\mathbf{x}^n - \mathbf{X}), \tag{9}$$

où

$$B = \left[\frac{\partial g_i}{\partial x_j}\right]\Big|_{\mathbf{X}}.$$

Ainsi, les approximations se trouvant dans un petit voisinage de la solution, les erreurs sur les itérés du processus (2) (et sur ceux de (7) et (8)) sont régies par des lois sensiblement identiques à celles gouvernant les erreurs commises dans les méthodes itératives de résolution des systèmes linéaires. On montre par conséquent que pour une approximation initiale suffisamment bonne, le processus (2) converge dans les mêmes conditions imposées à la matrice B que la méthode itérative simple. La validité de (9) permet également d'accélérer la convergence.

Examinons le cas $m = 1$ et construisons l'analogue du procédé δ^2. Etant donné x^n, introduisons les notations $x^{n,1} = g(x^n)$, $x^{n,2} = g(x^{n,1})$. Conformément à (9),

$$x^{n,1} - X \approx g'(X)(x^n - X),$$
$$x^{n,2} - X \approx g'(X)(x^{n,1} - X).$$

Il en découle

$$g'(X) \approx \frac{x^{n,2} - x^{n,1}}{x^{n,1} - x^n},$$

$$X \approx \frac{x^{n,2} - g'(X)\,x^{n,1}}{1 - g'(X)} \approx \frac{x^{n,2} - \dfrac{x^{n,2} - x^{n,1}}{x^{n,1} - x^n}\,x^{n,1}}{1 - \dfrac{x^{n,2} - x^{n,1}}{x^{n,1} - x^n}} = \frac{x^{n,2}x^n - (x^{n,1})^2}{x^{n,2} - 2x^{n,1} + x^n}.$$

Prenons pour l'approximation suivante

$$x^{n+1} = \frac{x^{n,2}x^n - (x^{n,1})^2}{x^{n,2} - 2x^{n,1} + x^n} = \frac{x^n g(g(x^n)) - (g(x^n))^2}{g(g(x^n)) - 2g(x^n) + x^n}. \qquad (10)$$

Pour caractériser les méthodes de résolution d'équations on introduit la notion *d'ordre de la méthode*. On dit qu'une méthode est d'ordre k s'il existe $c_1 > 0$, $c_2 < \infty$ tels que

$$\rho(\mathbf{x}^{n+1}, \mathbf{X}) \leqslant c_2 (\rho(\mathbf{x}^n, \mathbf{X}))^k$$

sous l'hypothèse $\rho(\mathbf{x}^n, \mathbf{X}) \leqslant c_1$. Il est clair que plus k est grand et plus vite converge le processus itératif pour $\rho(\mathbf{x}^n, \mathbf{X})$ faibles. Les calculateurs recourent le plus volontiers aux méthodes du premier ordre (p. ex. la méthode itérative simple) et du deuxième (p. ex. la méthode de Newton qui sera examinée au paragraphe suivant).

PROBLÈME. Démontrer que pour $g(x)$ régulière la méthode (10) est du deuxième ordre.

NOTE. Il existe une autre définition de l'ordre d'une méthode: on dit de la méthode de résolution du système $\mathbf{F}(\mathbf{x}) = 0$ qu'elle est d'ordre k si sa réalisation suppose le calcul des dérivées d'ordre $k - 1$ inclusivement de la fonction f_i.

Dans le § 12 du chapitre précédent nous avons traité des méthodes itératives de résolution des systèmes linéaires à l'aide d'opérateurs spectralement équivalents. Des procédés analogues existent pour les systèmes non linéaires. On choisit un opérateur $G(\mathbf{x})$ tel que $\mathbf{x} = 0$ soit l'unique solution de $G(\mathbf{x}) = 0$. Les approximations $\mathbf{x}^{n+1}$ de la solution du système $\mathbf{F}(\mathbf{x}) = 0$ se déterminent à partir de la relation

$$G(\mathbf{x}^{n+1} - \mathbf{x}^n) = F(\mathbf{x}^n). \qquad (11)$$

Le cas le plus répandu est celui où G est un opérateur linéaire. La relation (11) s'écrit alors

$$A(\mathbf{x}^{n+1} - \mathbf{x}^n) = F(\mathbf{x}^n); \tag{12}$$

si les vecteurs $\mathbf{x}$ sont de dimension m finie, A est une matrice. On prend souvent un G qui soit fonction de n et de l'approximation $\mathbf{x}^n$. Le schéma (12) joue alors pour la méthode de Newton ci-dessous.

§ 2. Méthode de Newton pour des systèmes d'équations non linéaires

Si l'on connaît une approximation initiale suffisamment bonne de la solution de

$$\mathbf{F}(\mathbf{x}) = 0, \tag{1}$$

on améliore efficacement la précision par la méthode de Newton.

Sans compliquer les raisonnements, bornons-nous à un cas plus général, à savoir la résolution d'une équation fonctionnelle non linéaire.

Soit $\mathbf{F}(\mathbf{x})$ un opérateur qui applique un espace vectoriel normé $\mathscr{X}$ sur un espace vectoriel normé $\mathscr{Y}$ qui se confond peut-être avec $\mathscr{X}$. Notons $\| \ \|_{\mathscr{X}}$ et $\| \ \|_{\mathscr{Y}}$ leurs normes respectives. L'opérateur linéaire P de $\mathscr{X}$ dans $\mathscr{Y}$ s'appelle *dérivée de l'opérateur* $\mathbf{F}(\mathbf{x})$ *au point* $\mathbf{x}$ si

$$\| \mathbf{F}(\mathbf{x}+\eta) - \mathbf{F}(\mathbf{x}) - P\eta \|_{\mathscr{Y}} = o(\| \eta \|_{\mathscr{X}}) \tag{2}$$

pour $\| \eta \|_{\mathscr{X}} \to 0$.

Dans la suite nous désignerons cet opérateur P par $F'(\mathbf{x})$. Soit, par exemple,

$$\mathbf{x} = (x_1, \ \ldots, \ x_m)^T, \quad \mathbf{F} = (f_1, \ \ldots, \ f_m)^T.$$

Si les fonctions f_i sont continûment dérivables au voisinage du point $\mathbf{x}$ donné, il vient

$$f_i(x_1 + \eta_1, \ \ldots, \ x_m + \eta_m) =$$

$$= f_i(x_1, \ \ldots, \ x_m) + \sum_{j=1}^{m} \frac{\partial f_i(x_1, \ \ldots, \ x_m)}{\partial x_j} \eta_j + o(\| \eta \|).$$

L'ensemble de ces relations se met sous la forme (2) à condition de prendre pour P l'opérateur de multiplication à gauche par la matrice

$$F'(\mathbf{x}) = \left[\frac{\partial f_i}{\partial x_j} \right].$$

Pour $m = 1$, cas le plus simple, P est l'opérateur de multiplication par la dérivée f'_x.

Soient X solution de $F(X) = 0$ et x^n une approximation de X. Dans l'hypothèse que F' existe, on a, en conformité avec (2),

$$\| F(X-)F(x^n) - F'(x^n)(X-x^n) \|_{\mathcal{Y}} = o(\| X-x^n \|_{\mathcal{X}}). \quad (3)$$

Si la quantité $\| X-x^n \|_{\mathcal{X}}$ est petite, on écrit l'égalité approchée

$$F(x^n) + F'(x^n)(X-x^n) \approx F(X).$$

Comme $F(X) = 0$, on a

$$F(x^n) + F'(x^n)(X-x^n) \approx 0.$$

Prenons pour approximation suivante la solution (si elle existe) de l'équation

$$F(x^n) + F'(x^n)(x^{n+1}-x^n) = 0$$

qui est de la forme (1.12). En supposant F' inversible, cette solution s'écrit

$$x^{n+1} = x^n - (F'(x^n))^{-1} F(x^n). \quad (4)$$

Ce processus itératif porte le nom de *méthode de Newton.*

Soit $\Omega_a = \{x : \| x-X \| < a\}$. Supposons qu'on a pour certains a, a_1, a_2, $0 < a$, $0 \leqslant a_1$, $a_2 < \infty$,

1. $$\| (F'(x))^{-1} \| \leqslant a_1 \quad \text{avec } x \in \Omega_a. \quad (5)$$

2. $$\| F(U_1 - F(U_2) - F'(U_2)(U_1 - U_2 \|_{\mathcal{Y}} \leqslant a_2 \| U_2 - U_1 \|^2_{\mathcal{X}} \quad (6)$$

avec U_1, $U_2 \in \Omega_a$. Notons $c = a_1 \cdot a_2$, $b = \min(a, c^{-1})$.

THÉORÈME (sur la convergence de la méthode de Newton). *Sous l'hypothèse* 1, 2 *et* $x^0 \in \Omega_b$, *le processus itératif de Newton* (4) *converge et l'erreur commise est évaluée par*

$$\| x^n - X \|_{\mathcal{X}} \leqslant c^{-1} (c \| x^0 - X \|_{\mathcal{X}})^{2^n}. \quad (7)$$

NOTE. Si, dans l'exemple ci-dessus, les fonctions f_i ont leurs dérivées secondes bornées, on a, en vertu de la formule de Taylor,

$$f_i(y) = f_i(x) + \sum_{j=1}^{n} \frac{\partial f_i(x_1, \ldots, x_n)}{\partial x_j} (y_j - x_j) + O(\| y-x \|^2),$$

et la condition 2 est donc remplie.

DÉMONSTRATION. Soit $x^0 \in \Omega_b$. Démontrons par récurrence sur n que tous les x^n appartiennent à Ω_b. Supposons que c'est vrai pour un certain n ; comme $b \leqslant a$, $x^n \in \Omega_a$. Substituons $U_1 = X$ et $U_2 = x^n$ dans (6), il vient

$$\| F(X) - F(x^n) - F'(x^n)(X-x^n) \|_{\mathcal{Y}} \leqslant a_2 \| x^n - X \|^2_{\mathcal{X}}.$$

Puisque $\mathbf{F}(\mathbf{x}^n) = -F'(\mathbf{x}^n)(\mathbf{x}^{n+1} - \mathbf{x}^n)$ et $\mathbf{F}(\mathbf{X}) = 0$, cette relation se récrit comme suit :

$$\| F'(\mathbf{x}^n)(\mathbf{x}^{n+1} - \mathbf{X}) \|_{\mathcal{Y}} \leqslant a_2 \| \mathbf{x}^n - \mathbf{X} \|_{\mathcal{X}}^2.$$

En se rappelant la condition (5), il en résulte

$$\| \mathbf{x}^{n+1} - \mathbf{X} \|_{\mathcal{X}} \leqslant c \| \mathbf{x}^n - \mathbf{X} \|_{\mathcal{X}}^2. \tag{8}$$

D'où

$$\| \mathbf{x}^{n+1} - \mathbf{X} \|_{\mathcal{X}} < cb^2 = (cb)\, b \leqslant b$$

et $\mathbf{x}^{n+1}$ appartient lui aussi à Ω_b. Ainsi, si $\mathbf{x}^0 \in \Omega_b$, il en est de tous les $\mathbf{x}^n$ qui vérifient donc l'inégalité (8).

Soit $q_n = c \| \mathbf{x}^n - \mathbf{X} \|_{\mathcal{X}}$. L'inégalité (8) multipliée par c donne $q_{n+1} \leqslant q_n^2$.

Par récurrence sur n démontrons la validité de

$$q_n \leqslant q_0^{2^n}.$$

Lorsque $n = 0$, cette inégalité est évidente. Supposons qu'elle l'est pour $n = k$ et obtenons

$$q_{k+1} \leqslant q_k^2 \leqslant (q_0^{2^k})^2 = q_0^{2^{k+1}}.$$

Donc, $q_n \leqslant q_0^{2^n}$ avec n quelconque, ce qui signifie que

$$c \| \mathbf{x}^n - \mathbf{X} \|_{\mathcal{X}} \leqslant (c \| \mathbf{x}^0 - \mathbf{X} \|_{\mathcal{X}})^{2^n};$$

il en découle (7). D'après la définition de c et b

$$c \| \mathbf{x}^0 - \mathbf{X} \|_{\mathcal{X}} < cb \leqslant 1,$$

c'est pourquoi $\mathbf{x}^n \to \mathbf{X}$. Le théorème est démontré.

L'estimation (7) de la vitesse de convergence de la méthode de Newton est une conséquence de l'inégalité (8) vraie pour toute méthode d'ordre 2. Ainsi, l'estimation (7) avec une certaine constante c a lieu dans toute méthode du deuxième ordre.

On a parfois, au lieu de (6), une estimation plus faible

$$\| \mathbf{F}(\mathbf{U}_1) - \mathbf{F}(\mathbf{U}_2) - F'(\mathbf{U}_2)(\mathbf{U}_1 - \mathbf{U}_2) \|_{\mathcal{Y}} \leqslant a_3 \| \mathbf{U}_2 - \mathbf{U}_1 \|_{\mathcal{X}}^p,$$

où $1 < p < 2$. (7) est alors remplacée par

$$\| \mathbf{x}^n - \mathbf{X} \|_{\mathcal{X}} \leqslant c_1^{-1} (c_1 \| \mathbf{x}^0 - \mathbf{X} \|_{\mathcal{X}})^{p^n}.$$

L'inversion de l'opérateur $F'(\mathbf{x}^n)$ s'avère souvent plus délicate que le calcul de $F(\mathbf{x}^n)$. C'est pourquoi la méthode de Newton se prête, dans de nombreux cas, à une modification suivante. En calculant on choisit ou bien on se donne d'avance une certaine suite

croissante de nombres $n_0 = 0, n_1, n_2, \ldots$ Si $n_k \leqslant n < n_{k+1}$, on effectue les itérations selon la formule

$$\mathbf{x}^{n+1} = \mathbf{x}^n - (F'(\mathbf{x}^{n_k}))^{-1} F(\mathbf{x}^n).$$

L'augmentation du nombre d'itérations qu'entraîne souvent cette variante de la méthode de Newton est contrebalancée par le faible coût d'un pas d'itération. On choisit la suite $\{n_k\}$ en prenant en considération tous ces facteurs.

Envisageons l'image géométrique de la méthode de Newton dans le cas de l'équation scalaire $f(x) = 0$. La formule de calcul (4) s'écrit alors

$$x^{n+1} = x^n - \frac{f(x^n)}{f'(x^n)}. \tag{9}$$

x^{n+1} s'obtient géométriquement comme l'abscisse du point d'intersection de l'axe des x et de la tangente à la courbe $y = f(x)$ au point $(x^n, f(x^n))$ (fig. 7.2.1). Si $f(x)$ est un polynôme du troisième degré et pour une approximation de départ médiocre il se peut que la suite $\{x^n\}$ ne converge pas vers la racine. Ainsi, dans le cas de la fig. 7.2.2, toutes les approximations paires coïncident avec a et les impaires avec b : la méthode cycle. Dans des problèmes plus compliqués l'allure réelle des x^n pour une mauvaise approximation initiale devient beaucoup plus embrouillée et se prête difficilement à une analyse quelconque.

Comparons la vitesse asymptotique de convergence de la méthode de Newton à celle de la méthode itérative simple. L'erreur résultant de celle-ci a été évaluée par

$$\| \mathbf{x}^n - \mathbf{X} \| \leqslant q^n \| \mathbf{x}^0 - \mathbf{X} \|, \quad q < 1.$$

Pour qu'elle devienne inférieure à ε, il suffit, d'après cette estimation, de prendre

$$n \geqslant \log_{q^{-1}} \frac{\| \mathbf{x}^0 - \mathbf{X} \|}{\varepsilon} \sim \log_{q^{-1}} \frac{1}{\varepsilon}.$$

Quant à la méthode de Newton, le deuxième membre de (7) est plus petit que ε si

$$n \geqslant -\log_2 \frac{\log_2 (c \| \mathbf{x}^0 - \mathbf{X} \|)}{\log_2 (c\varepsilon)} \sim \log_2 \log_2 \frac{1}{\varepsilon}. \tag{10}$$

Ainsi, pour $\varepsilon \to 0$, la méthode de Newton exige moins d'itérations.

On remarquera que ce même procédé mis sous la forme (4) est une variante de la méthode itérative simple. L'examen de l'équation scalaire $f(x) = 0$ révèle une autre particularité de la méthode de Newton. La dérivée par rapport à x du deuxième membre de (9) $g(x) = x - \frac{f(x)}{f'(x)}$ est égale à $\frac{f(x)f''(x)}{(f'(x))^2}$. Donc $g'(X) = 0$ si $f'(X) \neq \neq 0$ et la fig. 7.1.1 devient la fig. 7.2.3.

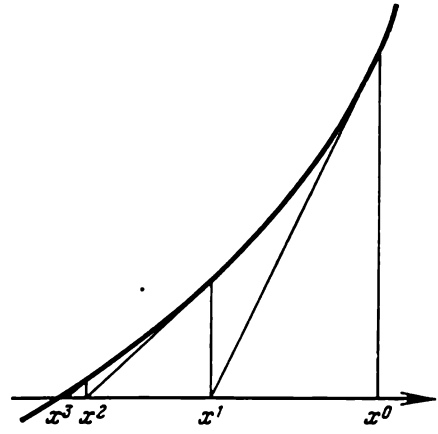


Fig. 7.2.1

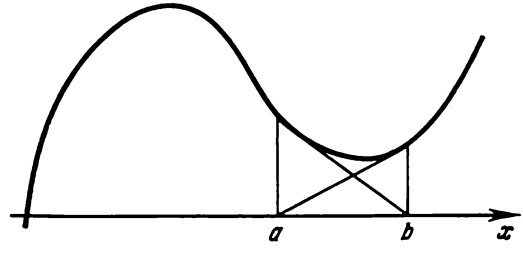


Fig. 7.2.2

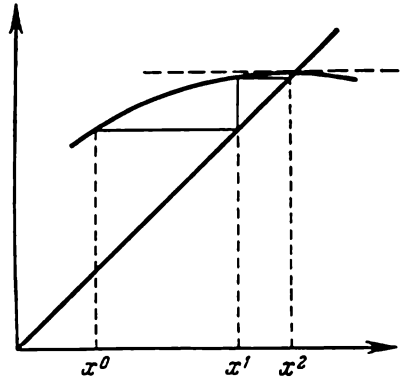


Fig. 7.2.3

La méthode de Newton s'avère fort intéressante pour l'extraction de racines p-ièmes avec p entier. Extraire la racine $\sqrt[p]{a}$, p entier, équivaut à résoudre l'équation $x^p - a = 0$. La formule de calcul de la méthode de Newton devient alors

$$x_{n+1} = \frac{p-1}{p} x_n + \frac{a}{p x_n^{p-1}}.$$

Pour approximation initiale de $\sqrt[p]{a}$ on peut prendre la valeur d'un polynôme approchant cette fonction.

PROBLÈME. Soit l'algorithme de calcul de $\sqrt{a}$ pour $1 \leqslant a \leqslant 4$, x_0 étant supposée égale à la valeur du polynôme de meilleure approximation uniforme de $\sqrt{a}$ sur $[1, 4]$: $x_0 = p_1(a) = \frac{17}{24} + \frac{a}{3}$. Vérifier que $|x_4 - \sqrt{a}| \leqslant 0{,}5 \cdot 10^{-25}$.

§ 3. Autres méthodes de résolution d'une équation

Si l'on possède une bonne approximation de la racine X de l'équation $f(x) = 0$, on emploie, à côté de la méthode de Newton, la *méthode de la sécante*.

Considérons sa variante la plus simple. On fixe, dans le processus itératif, un certain point x^0. L'itéré x^{n+1} s'obtient comme abscisse

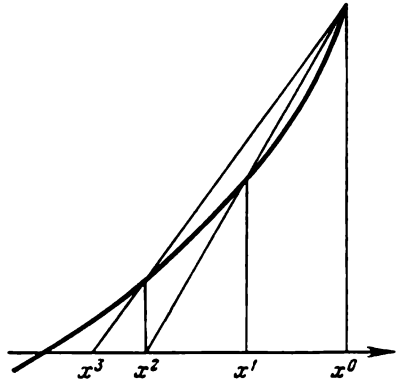


Fig. 7.3.1

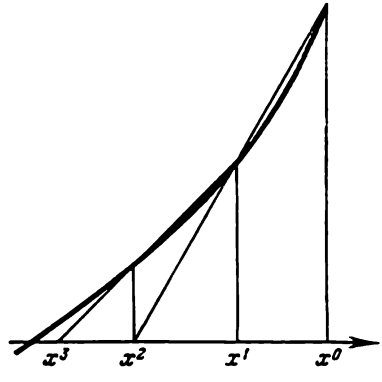


Fig. 7.3.2

du point d'intersection de la droite passant par $(x^0, f(x^0))$ et $(x^n, f(x^n))$ avec l'axe des x (fig. 7.3.1).

La méthode est plus efficace si l'on prend pour x^{n+1} l'abscisse du point en lequel l'axe des x coupe la droite passant par $(x^{n-1}, f(x^{n-1}))$ et $(x^n, f(x^n))$ (fig. 7.3.2). Cette droite a pour équation

$$y_n(x) = f(x^n) + (x - x^n) \frac{f(x^n) - f(x^{n-1})}{x^n - x^{n-1}}.$$

La condition $y_n(x^{n+1}) = 0$ fournit

$$x^{n+1} = x^n - \frac{f(x^n)(x^n - x^{n-1})}{f(x^n) - f(x^{n-1})}. \tag{1}$$

L'évaluation du nombre requis d'itérations se fait moyennant la relation qu'il faut établir entre les erreurs $r_n = x^n - X$ sur les approximations consécutives. On a l'égalité (développement de Taylor)

$$0 = f(X) = f(x^n) + f'(x^n)(X - x^n) + \frac{f''(x^n)}{2}(X - x^n)^2 + O((X - x^n)^3).$$

D'où

$$f(x^n) = f'(x^n) r_n - \frac{f''(x^n)}{2} r_n^2 + O(r_n^3).$$

De même

$$f(x^{n-1}) - f(x^n) = f'(x^n)(x^{n-1} - x^n) + \frac{f''(x^n)}{2}(x^{n-1} - x^n)^2 + O((x^{n-1} - x^n)^3).$$

Il en résulte

$$f(x^n) - f(x^{n-1}) = f'(x^n)(r_n - r_{n-1}) - \frac{f''(x^n)}{2}(r_n - r_{n-1})^2 + O((r_n - r_{n-1})^3).$$

En retranchant X de (1) il vient

$$r_{n+1} = r_n - \frac{\left(f'(x^n) r_n - \dfrac{f''(x^n)}{2} r_n^2 + O(r_n^3)\right)(r_n - r_{n-1})}{f'(x^n)\left(r_n - r_{n-1}\right) - \dfrac{f''(x^n)}{2}(r_n - r_{n-1})^2 + O((r_n - r_{n-1})^3)}.$$

Supposons $f'(x) \neq 0$ et divisons le numérateur et le dénominateur par $f'(x^n)(r_n - r_{n-1})$; on a

$$r_{n+1} = r_n - \frac{r_n - \lambda r_n^2 + O(r_n^3)}{1 - \lambda(r_n - r_{n-1}) + O((r_n - r_{n-1})^2)}$$

avec

$$\lambda = \frac{f''(x^n)}{2f'(x^n)}.$$

Compte tenu de

$$\frac{1}{1-\varepsilon} = 1 + \varepsilon + O(\varepsilon^2),$$

il en découle

$$r_{n+1} = r_n - (r_n - \lambda r_n^2 + O(r_n^3))(1 + \lambda(r_n - r_{n-1}) + O((r_n - r_{n-1})^2))$$

ou encore

$$r_{n+1} = \lambda r_n r_{n-1} + O((|r_n| + |r_{n-1}|)^3). \tag{2}$$

Nous nous abstiendrons d'évaluer strictement la vitesse de convergence du processus en nous limitant à des raisonnements vraisemblables. La relation (2) implique tout naturellement que l'erreur r_n diminue assez rapidement au cours des itérations :

$$r_{n-1} \gg r_n \gg r_{n+1}; \tag{3}$$

on n'a pourtant aucune raison d'espérer que, pour $n \to \infty$, r_{n+1} sera d'un ordre de grandeur inférieur à celui de $r_n r_{n-1}$. Puisque $r_{n-1} \gg r_n$, on a $r_{n-1} r_n \gg r_n^2$; cette inégalité et (2) entraînent $r_{n+1} \gg r_n^2$. En diminuant n d'une unité, on obtient $r_n \gg r_{n-1}^2$; par conséquent, $r_{n-1}^2 \ll r_n r_{n-1}$. Cette dernière inégalité et (3) déterminent $r_n^2 \ll r_n r_{n-1}$. Aussi, pour n grand, (2) s'écrit

$$r_{n+1} \sim \lambda r_n r_{n-1}.$$

La substitution $|\lambda r_n| = R_n$ donne

$$R_{n+1} \sim R_n R_{n-1}.$$

En en prenant le logarithme il vient

$$\ln R_{n+1} \sim \ln R_n + \ln R_{n-1}.$$

On est fondé de supposer que $\ln R_n$ (n grand) se comporte comme la solution de l'équation aux différences finies

$$u_{n+1} = u_n + u_{n-1};$$

l'équation caractéristique correspondante

$$\mu^2 - \mu - 1 = 0$$

a les racines

$$\mu_1 = \frac{1+\sqrt{5}}{2}, \qquad \mu_2 = \frac{1-\sqrt{5}}{2} \quad \text{avec } |\mu_2| < 1.$$

Donc $u_n = c_1\mu_1^n + c_2\mu_2^n \sim c_1\mu_1^n$ pour n tendant vers l'infini. Sous l'hypothèse $\ln R_n \sim u_n$ on écrit l'égalité asymptotique

$$\log_2 |\ln R_n| \sim \log_2 |c_1\mu_1^n| \sim n \log_2 \mu_1.$$

Les inégalités $|r_n| \leqslant \varepsilon$, $R \leqslant |\lambda| \varepsilon$ et $|\ln R_n| \geqslant |\ln (|\lambda| \varepsilon)|$ sont équivalentes lorsque $|\lambda| \varepsilon \leqslant 1$. C'est pourquoi on a $\log_2 |\ln (|\lambda| \varepsilon)| \sim n \log_2 \mu_1$ où n est le plus petit nombre pour lequel est réalisée la précision ε démandée; et donc

$$n \sim \frac{1}{\log_2 \mu_1} \log_2 \log_2 \frac{1}{\varepsilon}.$$

La comparaison avec (2.10) montre que dans la méthode de la sécante le nombre nécessaire d'itérations est, pour ε petit, d'environ $\log_{\mu_1} 2 \approx 1{,}44$ fois supérieur à celui qu'exige la méthode de Newton. Si d_1 est le volume de calcul de la valeur d'une fonction en un point et d_2 le volume de calcul supplémentaire de la dérivée en ce point, le coût total de la première méthode est dans notre cas de l'ordre de $1{,}44 d_1 \log_2 \log_2 \frac{1}{\varepsilon}$ et celui de la deuxième de l'ordre de $(d_1 + d_2) \log_2 \log_2 \frac{1}{\varepsilon}$.

Examinons un problème type: soit une fonction concrète $f(x, \alpha_1, \ldots, \alpha_n)$; on demande d'élaborer un programme qui, d'après des valeurs données de $\alpha_1, \ldots, \alpha_n$, fournit la racine de l'équation $f(x, \alpha_1, \ldots, \alpha_n) = 0$. Il arrive souvent que les programmes de calcul des fonctions f et f_x possèdent de nombreux blocs en commun, ce qui fait que $d_2 < 0{,}44 d_1$. Conformément aux évaluations ci-dessus et en présence d'une bonne approximation au départ, on préférera la méthode de Newton à celle de la sécante.

Lorsqu'on a à résoudre le système

$$\mathbf{F}(\mathbf{x}) = \mathbf{0},$$

on fait appel à un analogue de la méthode de la sécante. Supposons définies les approximations $\mathbf{x}^{n-m}, \ldots, \mathbf{x}^n$ et connues les valeurs

$$f_i(\mathbf{x}^{n-m}), \ldots, f_i(\mathbf{x}^n).$$

Soit $y = L_i(\mathbf{x})$ l'équation du plan passant par les points

$$(\mathbf{x}^{n-m}, f_i(\mathbf{x}^{n-m})), \ldots, (\mathbf{x}^n, f_i(\mathbf{x}^n));$$

adoptons pour approximation suivante $\mathbf{x}^{n+1}$ la solution du système d'équations

$$L_i(\mathbf{x}) = 0, \quad i = 1, \ldots, m. \quad |$$

Lorsque $m = 1$, on reconnaît la méthode de la sécante que nous venons de considérer. Si $m > 1$, ce procédé n'est utilisé que rarement et cela dans les cas où la précision exigée est faible.

En fait, lorsque n tend vers l'infini, cette méthode se caractérise par un « aplatissement » du simplexe m-dimensionnel de sommets aux points $\mathbf{x}^{n-m}, \ldots, \mathbf{x}^n$; la conséquence en est une détérioration rapide du conditionnement du système $L_i(\mathbf{x}) = 0$. L'algorithme de calcul devient sensible à l'erreur de calcul et ne converge plus.

A part les modes opératoires décrits, il en existe de nombreux autres où la fonction $f(x)$ est approchée au voisinage de la racine par une fonction $g(x)$ pour laquelle l'équation $g(x) = 0$ se résout sous forme explicite. Ces méthodes ne démarrent cependant qu'à condition de posséder une approximation satisfaisante de la solution.

Pour obtenir celle-là on recourt parfois à la *méthode d'encadrement*. On détermine d'abord a^0 et b^0 tels que $f(x)$ soit continue sur $[a^0, b^0]$ et $f(a^0) f(b^0) < 0$. Dans le cas de l'équation $f(x, \alpha_1, \ldots, \alpha_n) = 0$ on réussit souvent à expliciter de pareils a_0, b_0: $a_0(\alpha_1, \ldots, \alpha_n)$, $b_0(\alpha_1, \ldots, \alpha_n)$; on les trouve également en formant un tableau des valeurs de $f(x)$ en des points aléatoires ou ordonnés. Une fois le segment $[a^0, b^0]$ trouvé, on choisit un certain point $c^0 \in (a^0, b^0)$. Après avoir comparé les signes de $f(a^0)$, $f(c^0)$ et $f(b^0)$ on prend pour $[a^1, b^1]$ celui des segments $[a^0, c^0]$, $[c^0, b^0]$ aux extrémités duquel $f(x)$ est de signes contraires. Ensuite, on choisit un point $c^1 \in (a^1, b^1)$, et ainsi de suite. On obtient finalement une suite de segments emboîtés $[a^0, b^0] \supset [a^1, b^1] \supset \supset \ldots$ tels que $f(a^n) f(b^n) < 0$.

On peut adopter la stratégie suivante pour choisir c^n. Au départ, on prend $c^n = (a^n + b^n)/2$; on calcule $f(c^n)$ et on établit si les points

$$(a^n, f(a^n)), \quad (c^n, f(c^n)), \quad (b^n, f(b^n)) \tag{4}$$

sont alignés.

S'il en est ainsi, alors

$$f(a^n) - 2f(c^n) + f(b^n) = 0;$$

aussi la condition d'appartenance de ces points à une même droite est

$$|f(a^n) - 2f(c^n) + f(b^n)| < \gamma |f(b^n) - f(a^n)|$$

avec γ choisi expérimentalement.

Si la condition est satisfaite, on prend immédiatement pour c^{n+1} le point d'intersection de la sécante passant par

$$(a^{n+1}, f(a^{n+1})), \quad (b^{n+1}, f(b^{n+1}))$$

et de l'axe des x; sinon, on pose

$$c^{n+1} = (a^{n+1} + b^{n+1})/2$$

et on voit si les points

$$(a^{n+1}, f(a^{n+1})), f(c^{n+1}, f(c^{n+1})), (b^{n+1}, f(b^{n+1}))$$

se placent sur une même droite, et ainsi de suite.

Dans certains cas on prendra toujours pour c^{n+1} l'abscisse du point de l'intervalle (a_n, b_n) en lequel la parabole passant par les points (4) coupe l'axe des x.

Un problème important est l'élaboration de méthodes efficaces de résolution d'équations de diverses classes typiques. Telle est en particulier la *méthode des paraboles* qui permet de trouver les racines du polynôme $P(z) = a_0 z^m + \ldots + a_m$ à coefficients quelconques (réels ou complexes). Etant données les approximations z_{n-2}, z_{n-1}, z_n de la racine, l'approximation suivante est recherchée comme suit. On construit un polynôme d'interpolation du deuxième degré qui coïncide avec $P(z)$ en z_{n-2}, z_{n-1}, z_n. On prend en qualité de z_{n+1} la racine de ce polynôme la plus proche de z_n. L'analyse de l'application de la méthode a conditionné une certaine modification de ce schéma dans les programmes standard. Bien que la convergence de ce procédé n'ait pas été démontrée pour un polynôme et des conditions initiales arbitraires, on ne connaît pas de cas où il ne converge pas ou converge lentement. En recherchant plusieurs racines d'un polynôme on procède à une *élimination* : après avoir trouvé une racine $\tilde{z}$ de $P(z)$ on cherche les racines du polynôme $\frac{P(z)}{z-\tilde{z}}$.

§ 4. Méthodes de descente

Le problème de minimisation d'une fonctionnelle se résout le plus souvent par une *méthode de descente*. On détermine, pour une approximation donnée, un sens de décroissance de la fonctionnelle et on déplace l'approximation dans ce sens. Si la grandeur du déplacement n'est pas très grande, la valeur de la fonctionnelle diminue nécessairement. Considérons plusieurs techniques de descente.

En étudiant la convergence de la méthode de Seidel pour le système d'équations $A\mathbf{x} = \mathbf{b}$, $A > 0$, nous avons décrit la méthode cyclique *de descente séquentielle* qui permettait de minimiser la fonction $\Phi(x_1, \ldots, x_m)$. On s'est donné une approximation $\Phi(x_1^0, \ldots, x_m^0)$, puis on a recherché une valeur $x_1 = x_1^1$ qui réalise $\inf_{x_1} \Phi(x_1, x_2^0, \ldots, x_m^0)$, ensuite on a cherché une valeur $x_2 = x_2^1$ qui réalise $\inf_{x_2} \Phi(x_1^1, x_2, x_3^0, \ldots x_m^0)$, etc.; le processus se répète cycliquement.

Soit $P_k\mathbf{x}$ l'approximation obtenue en se déplaçant à partir de $\mathbf{x}$ suivant la coordonnée x_k. En numérotant les approximations dans l'ordre selon lequel elles s'obtiennent par déplacement suivant les coordonnées successives, on écrit

$$\mathbf{x}^1 = P_1\mathbf{x}, \qquad \mathbf{x}^2 = P_2P_1\mathbf{x}, \ \ldots,$$
$$\mathbf{x}^m = P_m \ldots P_1\mathbf{x}, \quad \mathbf{x}^{m+1} = P_1P_m \ldots P_1\mathbf{x}, \ \ldots$$

La mise en œuvre de cette méthode implique la minimisation de fonctions d'une variable. Considérons le problème de minimisation de la fonction d'une variable $\Phi(x)$, $x = x^0$ étant l'approximation initiale du point de minimum. La solution exacte étant en général impossible, on procède comme suit: on prend des valeurs $\bar{x}^0$, $\bar{\bar{x}}^0$ et on construit une parabole $y = P_2(x)$ vérifiant

$$P_2(x^0) = \Phi(x^0),\ P_2(\bar{x}^0) = \Phi(\bar{x}^0),\ P_2(\bar{\bar{x}}^0) = \Phi(\bar{\bar{x}}^0).$$

L'abscisse x du point de minimum $P_2(x)$ est acceptée comme l'approximation x^1. La suite de points obtenus par le procédé décrit ne converge pas nécessairement vers le point d'extrémum cherché de $\Phi(x)$.

On comprendra mieux les divers cas qui peuvent se présenter à la lumière des exemples suivants. Soient x^n les approximations obtenues avec cet algorithme.

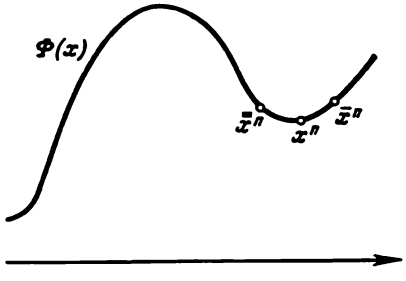


Fig. 7.4.1

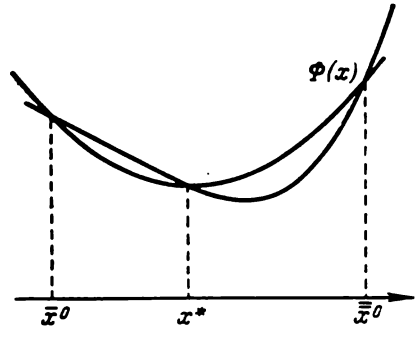


Fig. 7.4.2

Le choix à chaque pas d'itération des points $\bar{x}^n$, $\bar{\bar{x}}^n$ peut s'effectuer de diverses manières.

On les choisit de façon que

a) $$|\bar{x}^n - x^n| + |\bar{\bar{x}}^n - x^n| \longrightarrow 0 \quad \text{pour } n \longrightarrow \infty.$$

Il en résulte, par exemple, le cas de la fig. 7.4.1. Les points x^0, $\bar{x}^0$, $\bar{\bar{x}}^0$ se trouvent dans le voisinage immédiat du point d'extrémum local x^*, les points suivants x^n, $\bar{x}^n$, $\bar{\bar{x}}^n$ s'y trouvent également et $x^n \to x^*$; nous obtenons x^* au lieu du point d'extrémum absolu X qui nous intéresse.

b) $|\bar{x}^n - x^n| + |\bar{\bar{x}}^n - x^n| \not\to 0$; on choisit sans peine une fonction $\Phi(x)$ satisfaisant aux conditions ci-dessous (fig. 7.4.2):

1) il existe des points x^*, $\bar{x}^0$, $\bar{\bar{x}}^0$ tels que la parabole coïncidant en ces points avec $\Phi(x)$ admette un extrémum en x^*;

2) si $\bar{x}^n = \bar{x}^0$, $\bar{\bar{x}}^n = \bar{\bar{x}}^0$, $|x^n - x^*| \leqslant \eta$, $\eta > 0$, alors $x^{n+1} - x^* | \leqslant q\,|x^n - x^*|$, avec $q < 1$.

Avec une approximation au départ x^0 telle que $|x^0 - x^*| \leqslant \eta$, on peut s'illusionner sur la convergence du processus itératif vers x^* alors que celui-ci n'est même pas un point d'extrémum local.

Problème. Etablir une formule explicite pour une telle fonction $\Phi(x)$.

Même si l'on recherche à chaque pas l'extrémum absolu de $\Phi(x_1, \ldots, x_m)$ pour la coordonnée correspondante, on peut, dès

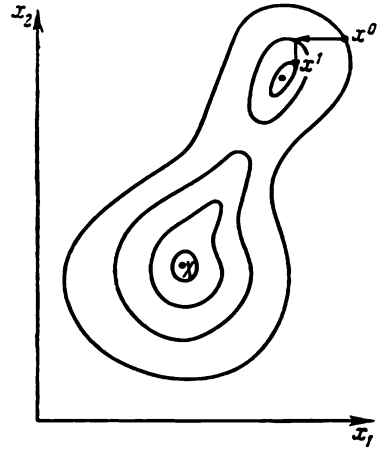


Fig. 7.4.3

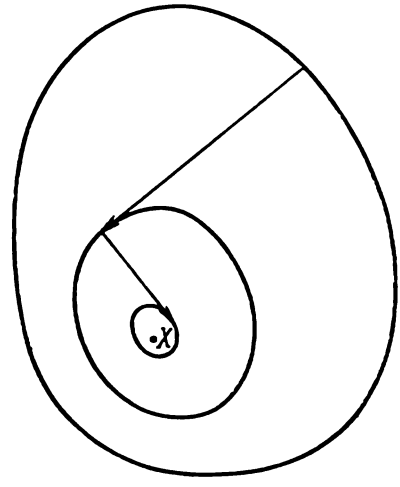


Fig. 7.4.4

$m = 2$, construire un processus itératif qui converge, pour une certaine approximation initiale, non pas vers le point d'extrémum absolu cherché, mais vers un point d'extrémum local. La fig. 7.4.3 visualise les courbes de niveau d'une telle fonction et les approximations obtenues.

Le déplacement n'est pas forcément cyclique. Si l'on sait, les résultats de calculs antérieurs aidant, que le déplacement dans un tel sens entraîne la plus grande diminution possible de $\Phi(x)$, on se déplace plus souvent suivant ces coordonnées.

Il se peut également qu'après avoir obtenu l'approximation $\mathbf{x}^n$ on choisisse, pour chaque n, un jeu de coordonnées

$$x_{i(n,1)}, \ldots, x_{i(n,q(n))}$$

et qu'on avance de façon indépendante suivant les coordonnées de ce groupe à partir de $\mathbf{x}^n$, i.e. on recherche les points $P_{i(n,k)}\mathbf{x}^n$. On calcule ensuite

$$\min_{1 \leqslant k \leqslant q(n)} \Phi(P_{i(n,k)}\mathbf{x}^n)$$

et le point $P_{i(n,k_0)}\mathbf{x}^n$ correspondant au minimum est pris pour $\mathbf{x}^{n+1}$.

Les procédés que nous venons de décrire sont les *méthodes de descente séquentielle optimale.*

Il y a parfois indétermination dans le choix du numéro de la coordonnée, ce qui permet de parler d'une *descente aléatoire.*

Une autre technique de descente est donnée par la *méthode de la plus grande pente* (ou *méthode du gradient*). L'approximation suivante est cherchée sous la forme

$$\mathbf{x}^{n+1} = \mathbf{x}^n - \Delta_n \operatorname{grad} \Phi(\mathbf{x}^n)$$

(fig. 7.4.4) et la valeur de Δ_n se définit par la condition

$$\min_{\Delta_n} \Phi(\mathbf{x}^n - \Delta_n \operatorname{grad} \Phi(\mathbf{x}^n)),$$

c'est-à-dire que l'algorithme consiste lui aussi à minimiser de proche en proche des fonctions d'une variable Δ_n.

Notons que la « descente séquentielle » et la « plus grande pente » ne nécessitent pas une résolution complète des problèmes de minimisation. Parfois, il est raisonnable, après avoir effectué un déplacement vers le minimum de la fonction auxiliaire d'une variable, de passer au pas suivant du processus itératif. Au voisinage de son point de minimum la fonction auxiliaire ne varie guère, c'est pourquoi une localisation trop exacte de ce point ne conduit pas à une diminution sensible des valeurs considérées.

S'agissant de la méthode de la plus grande pente, le volume de calcul exigé par la minimisation des fonctions auxiliaires d'une variable est un problème qui se résout compte tenu du coût relatif de recherche des valeurs de $\Phi(\mathbf{x})$ et de son gradient. Si le calcul des premières est essentiellement plus simple que celui du gradient, il n'est que légitime de mieux localiser le minimum à chaque pas d'itération ; sinon, on ferait mieux de ménager ses efforts.

Illustrons le choix d'une méthode sur l'exemple type suivant : soit un problème aux limites non linéaire pour un système d'équations différentielles ordinaires. Nous montrerons plus bas (chap. IX) que ce problème se ramène à résoudre le système non linéaire d'équations $\mathbf{F}(\mathbf{x}) = \mathbf{0}$:

$$f_i(\mathbf{x}) = f_i(x_1, \ldots, x_m) = 0, \quad i = 1, \ldots, m,$$

jouissant des propriétés suivantes.

La recherche d'une seule valeur $f_i(\mathbf{x})$ et celle de toutes les $f_i(\mathbf{x})$, $i = 1, \ldots, m$, à la fois au même point demandent un même volume de calcul (désignons-le par A). Il en est de même du calcul d'une $\frac{\partial f_i(\mathbf{x})}{\partial x_j}$ et de celui de toutes les $\frac{\partial f_i(\mathbf{x})}{\partial x_j}$, $i = 1, \ldots, m$; notons B l'effort dépensé dans ce cas. D'une façon générale ces coûts A et B sont très élevés et $B \gg A$.

Donc il est plus logique de calculer

$$\frac{\partial f_i(\mathbf{x})}{\partial x_j}, \quad j=1, \ldots, m,$$

à l'aide de la formule approchée

$$\frac{\partial f_i(x_1, \ldots, x_m)}{\partial x_j} \approx$$

$$\approx \frac{f_i(x_1, \ldots, x_{j-1}, x_j+\Delta, x_{j+1}, \ldots, x_m) - f_i(x_1, \ldots, x_{j-1}, x_j, x_{j+1}, \ldots, x_m)}{\Delta}. \quad (1)$$

La recherche de ces dérivées coûte alors $2A$, i.e. autant que celle de deux jeux de valeurs

$$f_i(x_1, \ldots, x_{j-1}, x_j, x_{j+1}, \ldots, x_m)$$

et

$$f_i(x_1, \ldots, x_{j-1}, x_j+\Delta, x_{j+1}, \ldots, x_m)$$

(abstraction faite de la soustraction et de la division par Δ).

En présence d'une bonne approximation de la solution on peut dans certains cas utiliser la méthode de Newton qui est rapidement convergente.

Le domaine de convergence de la méthode de Newton n'étant pas très vaste, il serait peut-être opportun, au moins au début du processus itératif, de ramener le problème à la minimisation d'une certaine fonctionnelle et de le résoudre par une technique de descente.

Examinons le cas simple de la fonctionnelle

$$\Phi(x) = \sum_{i=1}^{m} (\lambda_i f_i(x))^2,$$

les $\lambda_i = \text{const} \neq 0$ sont appelés *facteurs d'échelle*; en général on peut les choisir convenablement à partir des conditions de chaque problème concret.

Soit $\mathbf{x}^n = (x_1^n, \ldots, x_m^n)$ l'approximation obtenue à partir de laquelle on va avancer en x_1. Il faut alors calculer $f_i(\mathbf{x}^n)$, $i = 1, \ldots, m$ (si cela n'a pas été fait), et

$$f_i(\mathbf{x}^{n, \Delta}), \quad i=1, \ldots, m,$$

avec

$$\mathbf{x}^{n, \Delta} = (x_1^n+\Delta, x_2^n, \ldots, x_m^0);$$

ensuite, on trouve par (1) les approximations l_{i1} des dérivées $\frac{\partial f_i(\mathbf{x})}{\partial x_1}$.

Au voisinage du point $\mathbf{x}^n$, on a sur la droite passant par ce point parallèlement à l'axe des x_1:

$$f_i(\mathbf{x}) \approx f_i(\mathbf{x}^n) + l_{i1}(x_1 - x_1^n);$$

aussi la coordonnée x_1 de l'approximation suivante

$$\mathbf{x}^{n+1} = (x_1^{n+1}, x_2^n, \ldots, x_m^n)$$

se détermine à partir de la condition

$$\min_{x_1} \sum_{i=1}^{m} (\lambda_i (f_i(\mathbf{x}^n) + l_{i1}(x_1 - x_1^n)))^2.$$

La mise en œuvre d'un algorithme réel tiendra compte de plusieurs considérations complémentaires telles qu'un choix raisonnable de Δ par exemple: lorsque Δ est grand, la formule (1) donne lieu à une erreur d'approximation

importante, et, pour Δ petit, l'erreur dans cette formule s'avère également élevée puisque les erreurs sur f_i sont divisées par Δ.

Supposons que nous avons opté pour la méthode du gradient. Le calcul de grad $\Phi(x^n)$ ne se fait qu'en connaissant toutes les dérivées $\frac{\partial f_i(x^n)}{\partial x_j}$; en utilisant la formule (1) on aura un volume de calcul égal à $(m+1)A$; pour m grands, le travail $(m+1)A$ qu'exige le calcul de grad $\Phi(x)$ dépasse sensiblement le coût A de celui de $\Phi(x)$. Conséquence: il faut choisir plus scrupuleusement le point de minimum dans la méthode de la plus grande pente.

§ 5. Autres procédés de réduction de la dimension des problèmes

On trouve parfois intérêt à considérer le procédé formel suivant de réduction des problèmes multidimensionnels à des problèmes de dimension 1.

Soit à chercher le minimum Φ^* de la fonction

$$\Phi(x_1, \ldots, x_m)$$

dans le domaine

$$\begin{aligned} &\varphi_i(x_1, \ldots, x_m) \geqslant 0, \quad i = 1, \ldots, l, \\ &\psi_j(x_1, \ldots, x_m) = 0, \quad j = 1, \ldots, q. \end{aligned} \tag{1}$$

Il est possible d'écrire

$$\Phi^* = \min_{x_1, \ldots, x_m} \Phi(x_1, \ldots, x_m) = \min_{x_m} \Phi_m(x_m),$$

avec

$$\Phi_m(x_m) = \min_{x_1, \ldots, x_{m-1}} \Phi(x_1, \ldots, x_m);$$

chaque fois la fonction est minimisée sur son domaine de définition compte tenu des conditions (1). On a donc, au lieu du problème initial de minimisation d'une fonction de m variables, le problème de minimisation d'une fonction d'une variable dont chaque valeur se détermine en réalisant un minimum d'une fonction de $m-1$ variables. D'autre part, ramenons la minimisation de la fonction $\Phi_m(x_m)$ à la minimisation d'une fonction d'une variable telle que chacune de ses valeurs est déterminée en minimisant une fonction de $m-2$ variables, et ainsi de suite. On aboutit à la chaîne de relations

$$\begin{aligned} \Phi^* &= \min_{x_m} \Phi_m(x_m), \\ \Phi_m(x_m) &= \min_{x_{m-1}} \Phi_{m-1}(x_{m-1}, x_m), \\ &\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ \Phi_3(x_3, \ldots, x_m) &= \min_{x_2} \Phi_2(x_2, \ldots, x_m), \\ \Phi_2(x_2, \ldots, x_m) &= \min_{x_1} \Phi(x_1, \ldots, x_m). \end{aligned}$$

Cette méthode si simple à première vue est cependant très laborieuse. Supposons que chaque minimisation de fonctions d'une variable exige le calcul de s valeurs de la fonction à minimiser. La minimisation de $\Phi_m(x_m)$ exige qu'on minimise $\min_{x_{m-1}} \Phi_{m-1}(x_{m-1}, x_m)$, pour s valeurs du paramètre x_m, c est-à-dire s^2 calculs de $\Phi_{m-1}(x_{m-1}, x_m)$. Cela demande à son tour la recherche de s^3 valeurs de

$$\Phi_{m-2}(x_{m-2}, x_{m-1}, x_m),$$

etc. Finalement, il faut trouver s^m valeurs de la fonction Φ. Dès s et m moyens, disons $s = 10$, $m = 10$, le volume de calcul est inextricable. Par contre, avec un m petit, le procédé considéré admet plusieurs mises en œuvre intéressantes. On se donne par exemple une approximation au départ $(x_1^0, \ldots, x_m^0)$ et on réalise l'algorithme indiqué pour une valeur s pas très importante

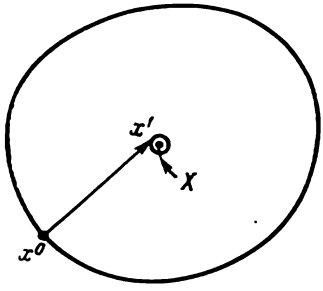


Fig. 7.5.1

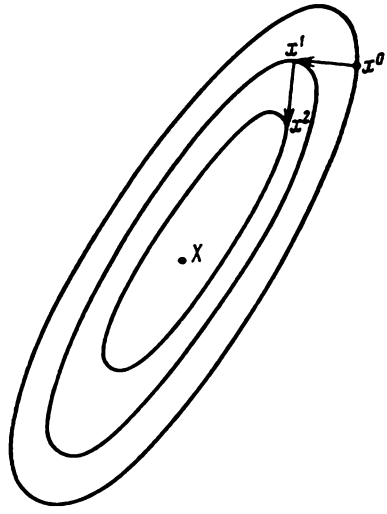


Fig. 7.5.2

($s = 3$ ou $s = 4$). Les valeurs de la fonction sont recherchées aux points d'un cube :

$$|x_1 - x_1^0|, \ldots, |x_m - x_m^0| \leqslant \Delta_0.$$

Le point de minimum $(x_1^1, \ldots, x_m^1)$ obtenu sert d'approximation suivante ; l'approximation $(x_1^2, \ldots, x_m^2)$ s'obtient de façon analogue, mais les valeurs de Φ sont calculées aux points du cube $|x_1 - x_1^1|, \ldots$ $\ldots, |x_m - x_m^1| \leqslant \Delta_1$, $\Delta_0 > \Delta_1 > \Delta_2 > \ldots$

Et voici une autre méthode de structure semblable au procédé précédent.

Si les courbes de niveau de la fonction Φ ressemblent à des sphères, les techniques de descente déterminent un rapide déplacement vers le minimum (fig. 7.5.1). Un cas plus typique est celui où

ces courbes ressemblent à des ellipsoïdes dont les demi-axes accusent une grande dispersion. Si l'on suit le gradient, on se déplace assez lentement dans la direction du minimum (fig. 7.5.2).

Supposons que les axes de ces « ellipsoïdes » se divisent naturellement en deux groupes dont le premier renferme m_1 axes relative-

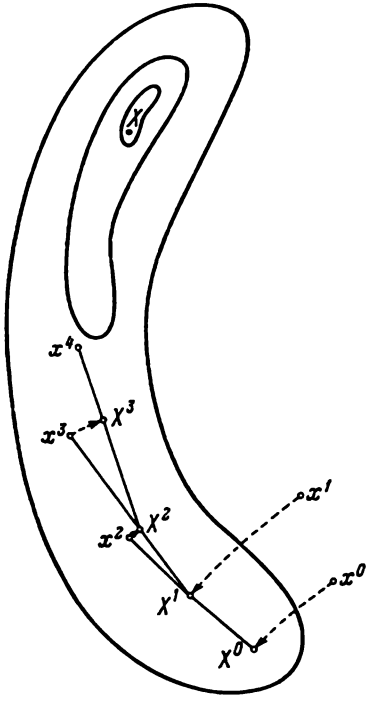


Fig. 7.5.3

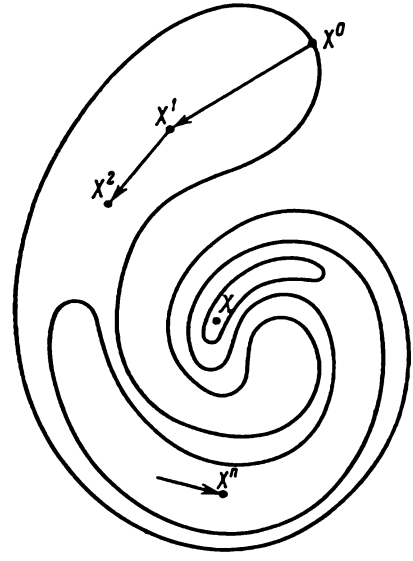


Fig. 7.5.4

ment petits et d'un même ordre de grandeur et le deuxième m_2 axes relativement grands et d'un même ordre de grandeur.

Dans certains problèmes de ce genre on utilise avec bonheur la *méthode* dite *des ravins*. On se donne des approximations $\mathbf{x}^0$ et $\mathbf{x}^1$ et on effectue à partir de chacune d'elles plusieurs pas par la technique de descente. Les approximations obtenues se notent $\mathbf{X}^0$ et $\mathbf{X}^1$. La fig. 7.5.3 révèle qu'elles appartiennent à une variété située au voisinage des axes du deuxième groupe. Le processus itératif consiste à rechercher de proche en proche $\mathbf{X}^0$, $\mathbf{X}^1$, . . . appartenant au voisinage de cette variété. Si $m_2 = 1$, $\mathbf{X}^{l+1}$ s'obtient comme suit. On mène une droite par $\mathbf{X}^{l-1}$ et $\mathbf{X}^l$ et on détermine l'approximation $\mathbf{x}^{l+1}$ du point de minimum de $\Phi(x)$ sur cette droite. Cette approxi-

mation se recherche donc sous la forme

$$\mathbf{x}^{l+1} = \mathbf{X}^l + \alpha (\mathbf{X}^l - \mathbf{X}^{l-1}).$$

Ensuite, avec au départ $\mathbf{x}^{l+1}$, effectuons plusieurs itérations, ce qui nous donne l'approximation $\mathbf{X}^{l+1}$, qui est située elle aussi dans le ravin. Si $m_2 > 1$, on recherche parfois $\mathbf{x}^{l+1}$ sous la forme

$$\mathbf{x}^{l+1} = \mathbf{X}^l + \alpha (\mathbf{X}^l - \mathbf{X}^{l-1}) + \beta \operatorname{grad} \Phi (\mathbf{X}^l).$$

La fig. 7.5.4 montre que la méthode décrite reste efficace dans certains cas où les courbes de niveau de Φ ont une forme plus compliquée.

La résolution du système d'équations

$$f_i (x_1, \ldots, x_m) = 0, \quad i = 1, \ldots, m, \tag{2}$$

se réduit formellement elle aussi à une résolution successive d'équations à une inconnue. Considérons le système d'équations

$$f_i (x_1, \ldots, x_m) = 0, \quad i = 2, \ldots, m, \tag{3}$$

par rapport à $x_2, \ldots, x_m$, et soit $x_2 (x_1), \ldots, x_m (x_1)$ sa solution. Portons les expressions de $x_2 (x_1), \ldots, x_m (x_1)$ dans la première équation (2); nous obtenons l'équation en x_1

$$F_1 (x_1) = f_1 (x_1, x_2 (x_1), \ldots, x_m (x_1)) = 0. \tag{4}$$

La recherche des valeurs $x_2 (x_1), \ldots, x_m (x_1)$ et la résolution de (4) peuvent se faire par un procédé numérique: on choisit une méthode de résolution de (4) d'après les valeurs de $F_1 (x_1)$. En résolvant (3), pour chaque x_1 cherché, on trouve les valeurs $x_2 (x_1), \ldots$ $\ldots, x_m (x_1)$ que l'on porte ensuite dans le deuxième membre de (4). La résolution du système (3) fait appel pour chaque valeur de x_1 au même mode opératoire.

Soit $x_3 (x_1, x_2), \ldots, x_m (x_1, x_2)$ la solution du système d'équations

$$f_i (x_1, \ldots, x_m) = 0, \quad i = 3, \ldots, m, \tag{5}$$

relativement aux inconnues $x_3, \ldots, x_m$. En portant cette solution dans la deuxième équation du système on obtient

$$F_2 (x_1, x_2) = f_2 (x_1, x_2, x_3 (x_1, x_2), \ldots, x_m (x_1, x_2)) = 0.$$

Elle se résout par rapport à x_2 pour tout x_1. Sa solution $x_2 (x_1)$ et $x_3 (x_1, x_2 (x_1)), \ldots, x_m (x_1, x_2 (x_1))$ forment la solution du système (3). On résout le système (5) pour chaque valeur de x_1 et x_2 en se ramenant à un système dont le nombre d'inconnues est inférieur d'une unité, et ainsi de suite.

Si la recherche de chaque équation auxiliaire à une inconnue exige s calculs de fonctions, cet algorithme coûte en tout $s + \ldots + s^m$ calculs des deuxièmes membres des équations du système.

De même que la méthode de minimisation décrite au début de ce paragraphe, l'algorithme en question est d'un emploi assez rare et il s'applique d'habitude dans le cas de m faible et pour une position relative compliquée des surfaces $f_i(x_1, \ldots, x_m) = 0$.

PROBLÈME. Que devient la méthode décrite pour un système (2) linéaire?

§ 6. Méthodes de stationnarisation

On résout très souvent les problèmes du type considéré par une *méthode de stationnarisation*, i.e. on construit un processus non stationnaire dont la solution tend à la limite, pour $t \to \infty$, vers la solution du problème initial. Soit un système d'équations différentielles

$$\frac{dx}{dt} + \operatorname{grad} \Phi = 0. \tag{1}$$

Le vecteur $\frac{dx}{dt}$ est proportionnel au gradient de Φ, i.e. orthogonal à ses courbes de niveau et dirigé dans le sens des Φ décroissants. Ainsi, lors du déplacement sur la trajectoire du système (1) la valeur de Φ ne croît pas. Cette affirmation est formellement légitimée par l'inégalité

$$\frac{d\Phi(x)}{dt} = \left(\operatorname{grad} \Phi, \frac{dx}{dt}\right) = -(\operatorname{grad} \Phi, \operatorname{grad} \Phi), \tag{2}$$

qui signifie que $\frac{d\Phi(x)}{dt}$ est inférieure à zéro partout sauf aux points stationnaires de la fonction $\Phi(x)$.

Un autre processus non stationnaire dont la solution tend à la limite, sous des hypothèses très générales, vers un point de minimum de $F(x)$ est gouverné par le système différentiel

$$\frac{d^2x}{dt^2} + \gamma \frac{dx}{dt} + \operatorname{grad} \Phi = 0, \quad \gamma > 0. \tag{3}$$

On a pour ce système

$$\frac{d}{dt}\left(\frac{1}{2}\left(\frac{dx}{dt}, \frac{dx}{dt}\right) + \Phi(x)\right) = \left(\frac{dx}{dt}, \frac{d^2x}{dt^2}\right) + \left(\operatorname{grad} \Phi, \frac{dx}{dt}\right) =$$

$$= -\gamma\left(\frac{dx}{dt}, \frac{dx}{dt}\right) < 0, \tag{4}$$

à condition que $\frac{dx}{dt} \neq 0$. Dans les deux cas, on peut identifier les fonctions $\Phi(x)$ et $\frac{1}{2}\left(\frac{dx}{dt}, \frac{dx}{dt}\right) + \Phi(x)$ à l'énergie d'un système matériel dont le mouvement est gouverné par les systèmes d'équations (1) et (3). Les relations (2) et (4) montrent que les processus non stationnaires considérés se caractérisent par une *dissipation* d'énergie.

Pour nous guider dans le choix de γ considérons un modèle élémentaire: x scalaire, $\Phi(x) = \frac{a^2x^2}{2}$; l'égalité (3) se récrit alors

$$x'' + \gamma x' + a^2 x = 0$$

et l'équation caractéristique correspondante

$$\lambda^2 + \gamma\lambda + a^2 = 0;$$

ses racines sont $\lambda_{1,2} = -\frac{\gamma}{2} \pm \sqrt{\frac{\gamma^2}{4} - a^2}$ et sa solution générale $c_1 \exp(\lambda_1 t) + c_2 \exp(\lambda_2 t)$ pour $\lambda_1 \neq \lambda_2$, i.e. $\gamma \neq 2a$.

Les solutions de l'équation analysée décroissent avec une vitesse déterminée par

$$\sigma(\gamma) = \max(\operatorname{Re}\lambda_1, \operatorname{Re}\lambda_2).$$

Si $\gamma \leqslant 2a$, on a $\frac{\gamma^2}{4} - a^2 \leqslant 0$, donc

$$\operatorname{Re}\lambda_1 = \operatorname{Re}\lambda_2 = \sigma(\gamma) = -\frac{\gamma}{2} \geqslant -a.$$

Si $\gamma > 2a$, λ_1 et λ_2 sont réels et évidemment

$$\operatorname{Re}\lambda_1 = -\frac{\gamma}{2} + \sqrt{\frac{\gamma^2}{4} - a^2} > \operatorname{Re}\lambda_2 = -\frac{\gamma}{2} - \sqrt{\frac{\gamma^2}{4} - a^2}.$$

Alors

$$\sigma(\gamma) = \operatorname{Re}\lambda_1 = -\frac{\gamma}{2} + \sqrt{\frac{\gamma^2}{4} - a^2} =$$

$$= -\frac{a^2}{\frac{\gamma}{2} + \sqrt{\frac{\gamma^2}{4} - a^2}} > -\frac{a^2}{\frac{\gamma}{2}} > -a.$$

Ainsi, le graphique de $\sigma(\gamma)$ est de la forme représentée sur la fig. 7.6.1 et $\min_{\gamma} \sigma(\gamma) = \sigma(2a) = -a$. Ce problème modèle autorise les conclusions suivantes.

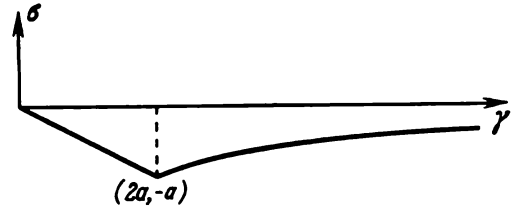


Fig. 7.6.1

1. Si le coefficient de frottement γ est insignifiant ($\gamma \ll 2a$ dans notre cas), la solution du système (3) tend lentement vers la

position d'équilibre ; ceci étant (vu la condition Im λ_1, Im $\lambda_2 \neq 0$), elle oscille près du point d'équilibre.

2. Si γ est fort ($\gamma \gg 2a$), la solution s'équilibre toujours lentement ; la cause en est l'impossibilité, pour le mouvement, de devenir suffisamment rapide lorsque le coefficient de frottement γ est important.

3. La valeur optimale de γ est située quelque part au milieu et dépend des propriétés de chaque fonction $\Phi(x)$ concrète.

La méthode de stationnaristion à l'aide du système (3) est quelquefois appelée *méthode de la boule pesante.*

Considérons, en effet, le mouvement d'un point matériel sur la surface $y = \Phi(x)$ dans le champ de pesanteur orienté dans le sens des y négatifs et supposons que le frottement est proportionnel à la vitesse et que le point ne peut quitter la surface. Son mouvement est alors décrit par le système d'équations

$$\frac{d^2x}{dt^2} + \gamma \frac{dx}{dt} + \frac{\operatorname{grad} \Phi}{1 + \| \operatorname{grad} \Phi \|^2} = 0.$$

Il est clair que la solution de ce système tend à la limite vers un certain point stationnaire de la fonction $\Phi(x)$. Au voisinage des points stationnaires $\| \operatorname{grad} \Phi \| \ll 1$ et notre système est proche du système (3).

La plupart des méthodes de stationnarisation connues sont gouvernées par des équations de la forme

$$A_0\left(x, \frac{dx}{dt}\right)\frac{dx}{dt} + A_1\left(x, \frac{dx}{dt}, \operatorname{grad} \Phi\right) = 0,$$

ou

$$B_0\left(x, \frac{dx}{dt}\right)\frac{d^2x}{dt^2} + B_1\left(x, \frac{dx}{dt}, \operatorname{grad} \Phi\right) = 0,$$

où

$$A_0(X, 0) \neq 0, \quad A_1(X, 0, 0) = 0,$$
$$B_0(X, 0) \neq 0, \quad B_1(X, 0, 0) = 0$$

et sont réalisées les conditions de dissipation qui garantissent la convergence vers le point d'extrémum. Formellement, on peut inverser les opérateurs A_0 et B_0 et transformer ces équations de façon que A_0 et B_0 soient identiques. Cependant, l'écriture initiale s'avère souvent plus commode.

On croirait que la construction de tels processus non stationnaires tendant à la limite vers la solution résout complètement le problème de recherche du minimum et qu'il ne reste « plus » qu'à intégrer le système obtenu d'équations différentielles à l'aide d'un programme standard approprié.

En réalité, ce problème n'est pas toujours résolu de façon satisfaisante par la réduction d'un problème stationnaire à un problème qui ne l'est pas. Il reste une question importante en suspens, celle du choix de la longueur des pas d'intégration. Supposons que la

stationnarisation a lieu avec la précision demandée pendant un intervalle de temps T. Si le pas d'intégration Δ est petit, les points obtenus avoisinent la trajectoire considérée et on peut espérer tomber dans un voisinage restreint du point de minimum. Cependant, le nombre de pas T/Δ peut s'avérer inadmissible (fig. 7.6.2). Si, au

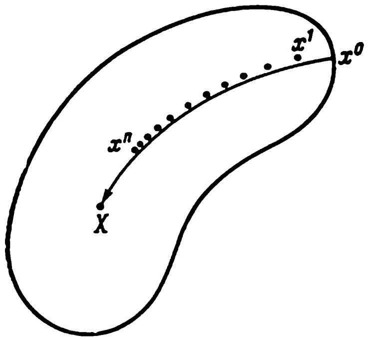


Fig. 7.6.2

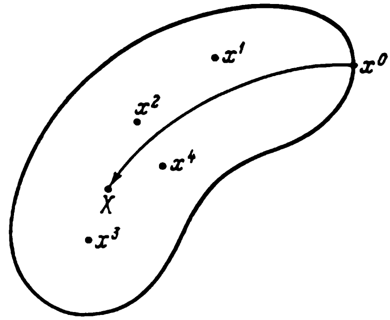


Fig. 7.6.3

contraire, on intègre avec un pas très important, on risque de voir les points s'écarter fortement de la solution considérée et ne jamais pénétrer dans le voisinage cherché du point de minimum (fig. 7.6.3).

Pour comprendre ce mécanisme adressons-nous à la fonctionnelle

$$\Phi^0(\mathbf{x}) = \frac{1}{2}(A\mathbf{x}, \mathbf{x}) - (\mathbf{b}, \mathbf{x}), \tag{5}$$

dont le minimum est réalisé, pour $A > 0$, sur la solution du système $A\mathbf{x} = \mathbf{b}$. Le système (1) s'écrit alors

$$\frac{d\mathbf{x}}{dt} + (A\mathbf{x} - \mathbf{b}) = 0.$$

En remplaçant la dérivée $\frac{d\mathbf{x}}{dt}$ par un rapport aux différences simple, il vient

$$\frac{\mathbf{x}(t_{n+1}) - \mathbf{x}(t_n)}{t_{n+1} - t_n} + A\mathbf{x}(t_n) - \mathbf{b} \approx 0;$$

la formule de calcul correspondante est

$$\mathbf{x}^{n+1} = \mathbf{x}^n - \Delta_n (A\mathbf{x}^n - \mathbf{b}), \quad \Delta_n = t_{n+1} - t_n.$$

Soit le cas $\Delta_n \equiv \Delta = \text{const} \neq 0$. On a alors

$$\mathbf{x}^{n+1} = \mathbf{x}^n - \Delta(A\mathbf{x}^n - \mathbf{b}).$$

Mettons cette relation sous la forme $\mathbf{x}^{n+1} = B\mathbf{x}^n + \mathbf{c}$, où $B = E - \Delta A$, $\mathbf{c} = \Delta\mathbf{b}$; soient $\lambda_1 \leqslant \ldots \leqslant \lambda_m$ les valeurs propres

de A, $\lambda_m/\lambda_1 \gg 1$. Les valeurs propres de la matrice B sont $1 - \Delta\lambda_i$. Si Δ est très petit, la valeur propre $1 - \Delta\lambda_i$ est proche de l'unité, et le nombre d'itérations est grand. Pour un Δ notable, la valeur propre $1 - \Delta\lambda_m$ a un module supérieur à 1 et le processus itératif diverge. Plus haut nous avons défini la valeur optimale $\Delta = 2/(M + \mu)$ du pas dans un processus pareil pour une matrice A dont les valeurs propres sont comprises entre des bornes M, μ connues. Si la valeur Δ_n est définie à chaque pas par la condition $\min_{\Delta_n} \Phi^0 (x^n - \Delta_n \operatorname{grad} \Phi^0 (x^n))$, la méthode de stationnarisation coïncide dans le cas considéré avec le procédé de la plus grande pente (pour les particularités de l'intégration de tels systèmes, voir § 11, chap. VIII).

Donnons-nous un pas constant $\Delta_n = t_{n+1} - t_n \equiv \Delta$ et écrivons l'une des approximations les plus élémentaires de l'équation (3):

$$\frac{x^{n+1} - 2x^n + x^{n-1}}{\Delta^2} + \gamma \frac{x^{n+1} - x^{n-1}}{2\Delta} + \operatorname{grad} \Phi (x^n) = 0.$$

En se plaçant dans le cas de la fonctionnelle (5) cette égalité devient

$$x^{n+1} = x^n + \frac{1 - \frac{\gamma\Delta}{2}}{1 + \frac{\gamma\Delta}{2}} (x^n - x^{n-1}) - \frac{\Delta^2}{1 + \frac{\gamma\Delta}{2}} (Ax^n - b).$$

En choisissant Δ et γ de façon appropriée, cette relation se met sous la forme (6.7.21) et on obtient un processus itératif qui ne diffère guère du processus linéaire optimal (6.7.16). Nous avons vu (§ 7, chap. VI) que celui-ci permet d'obtenir la solution avec une précision donnée au bout d'un nombre d'itérations sensiblement moins grand que celui qu'exige la méthode itérative simple. C'est pourquoi la stationnarisation par intégration numérique du système (3) est généralement plus efficace que celle correspondant à l'intégration du système (1).

A présent, donnons-nous une suite $t_0, t_1, \ldots$ d'intervalles de temps et remplaçons les dérivées de (3) par des différences divisées. Avec l'hypothèse de γ dépendant de t_n, il vient

$$2 \frac{\frac{x^{n+1} - x^n}{t_{n+1} - t_n} - \frac{x^n - x^{n-1}}{t_n - t_{n-1}}}{t_{n+1} - t_{n-1}} + \gamma_n \frac{x^{n+1} - x^{n-1}}{t_{n+1} - t_{n-1}} + (Ax^n - b) = 0.$$

A condition de choisir convenablement les paramètres γ_n et t_n, ces relations peuvent décrire le processus linéaire optimal (6.7.15) de même que la méthode du gradient conjugué.

L'analogue de cette dernière méthode dans le cas de la recherche du minimum d'une fonctionnelle de forme générale $\Phi(x)$ peut être

la technique suivante: les approximations consécutives sont cherchées comme

$$\mathbf{x}^{n+1} = \mathbf{x}^n + \alpha_n (\mathbf{x}^n - \mathbf{x}^{n-1}) + \beta_n \operatorname{grad} \Phi (\mathbf{x}^n)$$

avec α_n et β_n définis à partir de la condition $\min\limits_{\alpha_n, \beta_n} \Phi (\mathbf{x}^n)$.

Cette correspondance entre les méthodes de résolution des problèmes stationnaires par stationnarisation et les procédés par itérations ordinaires permet de découvrir des méthodes itératives et des processus de stationnarisation inédits.

Considérons, par exemple, les formules de Newton

$$\mathbf{x}^{n+1} - \mathbf{x}^n = -(F'(\mathbf{x}^n))^{-1} \mathbf{F}(\mathbf{x}^n) \tag{6}$$

de résolution du système non linéaire $\mathbf{F}(\mathbf{x}) = \mathbf{0}$. Elles s'interprètent comme suit: introduisons un t continu et prenons $\mathbf{x}^n$ pour valeurs d'une certaine fonction aux instants $t_n = n$. La relation (6) se récrit alors:

$$\frac{\mathbf{x}^{n+1} - \mathbf{x}^n}{t_{n+1} - t_n} = -(F'(\mathbf{x}^n))^{-1} \mathbf{F}(\mathbf{x}^n),$$

et est estimée être le résultat de l'approximation du système

$$\frac{d\mathbf{x}}{dt} = -(F'(\mathbf{x}))^{-1} \mathbf{F}(\mathbf{x}). \tag{7}$$

La solution du problème initial est un point stationnaire de ce système. Pour $\mathbf{x} = \mathbf{X} + \boldsymbol{\eta}$ on a

$$(F'(\mathbf{x}))^{-1} = (F'(\mathbf{X}))^{-1} + O(\|\boldsymbol{\eta}\|),$$

$$\mathbf{F}(\mathbf{x}) = \mathbf{F}(\mathbf{X}) + F'(\mathbf{X})\,\boldsymbol{\eta} + O(\|\boldsymbol{\eta}\|^2) = F'(\mathbf{X})\,\boldsymbol{\eta} + O(\|\boldsymbol{\eta}\|^2),$$

donc

$$\frac{d\boldsymbol{\eta}}{dt} = \frac{d\mathbf{x}}{dt} = -((F'(\mathbf{X}))^{-1} + O(\|\boldsymbol{\eta}\|))(F'(\mathbf{X})\,\boldsymbol{\eta} + O(\|\boldsymbol{\eta}\|^2)) =$$

$$= -\boldsymbol{\eta} + O(\|\boldsymbol{\eta}\|^2).$$

Il en découle que $\mathbf{x} = \mathbf{X}$ est une solution asymptotiquement stable de (7).

En substituant à la dérivée $\frac{d\mathbf{x}}{dt}$ le rapport aux différences relativement aux valeurs de la fonction en certains points $t_0 = 0$, t_1, . . ., il vient

$$\frac{\mathbf{x}^{n+1} - \mathbf{x}^n}{\Delta_n} = -(F'(\mathbf{x}^n))^{-1} \mathbf{F}(\mathbf{x}^n), \quad \Delta_n = t_{n+1} - t_n,$$

ou encore

$$\mathbf{x}^{n+1} = \mathbf{x}^n - \Delta_n (F'(\mathbf{x}^n))^{-1} F(\mathbf{x}^n). \tag{8}$$

La construction d'un processus non stationnaire correspondant à la méthode de Newton a donc eu pour résultat le processus itératif (8) d'une forme plus générale.

Cet exemple montre que le passage d'un algorithme d'itération au processus non stationnaire correspondant a beaucoup de commun avec la construction de la fermeture d'un algorithme de calcul (la notion de fermeture d'un algorithme sera introduite au § 4, chap. IX). Nous n'avons démontré la convergence de la méthode de Newton que pour une approximation initiale assez bonne de la solution. Pour élargir le domaine de convergence on utilise parfois la variante suivante de la méthode de Newton.

On se donne une fonctionnelle $\Phi(x)$, disons $\| F(x) \|^2$, dont la borne inférieure est réalisée par la solution du problème. On recherche les approximations successives sous la forme (8), Δ_n étant défini à partir de la condition

$$\min_{\Delta_n} \Phi(x^n - \Delta_n (F'(x^n))^{-1} F(x^n)).$$

Il s'agit alors de trouver la valeur de Δ_n réalisant ce minimum.

Voici l'une des procédures répandues. On se donne $\lambda \in (0, 1)$, $\theta \in (0, 1]$ et $l > 0$ entier, on calcule de proche en proche, pour chaque n,

$$x^{n+1, i} = x^n - \lambda^i (F'(x^n))^{-1} F(x^n), \quad i = 0, \ldots, l,$$

et $q^{n+1, i} = \Phi(x^{n+1, i})$. Si, avec un $k \leqslant l$, on a $q^{n+1, k} \leqslant \theta \Phi(x^n)$, on arrête le calcul et on pose $x^{n+1} = x^{n+1, k}$; dans d'autres procédures, on trouve $\min_{0<i\leqslant l} q^{n+1, i} = q^{n+1, m}$ et on pose $x^{n+1} = x^{n+1, m}$.

Si

$$q^{n+1, 0}, \ldots, q^{n+1, l} > \theta \Phi(x^n),$$

il en résulte, selon le programme,

1) un passage momentané à une autre méthode,
2) l'arrêt,
3) une variation des paramètres l, λ, θ.

Notons en guise de conclusion que les méthodes de stationnarisation s'appliquent également lorsqu'il s'agit de minimiser une fonction dans des domaines avec contraintes. On complète alors (1), (3) par des équations régissant la trajectoire du point tombant sur la frontière.

§ 7. Que faut-il optimiser ?

Nous venons d'examiner plusieurs procédés de minimisation de fonctions et de résolution de systèmes d'équations non linéaires. Bien que la liste des méthodes soit à peine entamée, passons à d'autres questions aussi importantes.

De nombreux problèmes de gestion industrielle et agricole, de transports, de répartition de ressources, etc., se ramènent, nous l'avons déjà dit, à minimiser des fonctionnelles (fonctions). Dans ces cas il est d'usage d'appeler ces problèmes de minimisation *problèmes d'optimisation* étant donné que le but qu'on s'assigne en premier lieu est de réaliser un régime de travail optimal. La fonction à minimiser est alors appelée *objectif* ou *fonction économique*.

La résolution des problèmes d'optimisation comprend l'élaboration du modèle mathématique du phénomène, la définition de l'objectif et des paramètres les plus importants à optimiser, la minimisation directe d'une fonction (dépendant en général d'un grand nombre de variables), la mise en pratique des résultats de l'étude.

Il est naturel que les deux premières et la dernière question soient examinées par les mathématiciens de concert avec les représentants de la branche intéressée.

Considérons un exemple réel de choix d'un modèle. Pour produire du sucre on doit cultiver la betterave sucrière, la transporter jusqu'aux usines et lui faire subir un traitement approprié. Les mathématiciens qui se sont proposé d'optimiser la production en question ont entrepris en premier lieu de minimiser les frais de transport, problème qui a été résolu avec succès à l'aide des machines à calculer. Les responsables ont cependant décliné le programme ainsi obtenu, car, selon eux, il ferait empirer dans une certaine mesure les conditions de transport d'autres cultures agricoles et compliquerait l'organisation du traitement: certaines sucreries seront surchargées de travail tandis que d'autres ne fonctionneront pas à plein rendement.

Voyons le bien fondé de ces objections.

Le prix de revient du sucre se compose (les chiffres sont conventionnels):
du coût de la betterave à sucre (80 %),
des frais de transport (5 %),
des frais de transformation (15 %).

La minimisation des frais de transport réduit ces derniers de 10 %, soit 0,5 % du prix de revient du sucre. Par contre, ce remaniement demande de gros efforts et l'on ne sait pas trop si d'autres pertes plus grandes encore ne viendront pas se greffer à cause de certains facteurs qui auront été négligés.

La structure ci-dessus du prix de revient montre que l'optimisation des frais de transport et de transformation ne permet pas une augmentation sensible de l'efficacité de la production. En approfondissant le problème on a constaté que le rendement en sucre par tonne de betterave est proportionnel à la teneur en sucre de la plante; or, le paramètre teneur en sucre varie selon une loi déterminée durant la maturation et diminue avec le stockage. En particulier, une récolte tardive est cause de pertes importantes du fait que de grandes quantités de betterave arrivent simultanément dans les sucreries qui les mettent pour longtemps en magasin. On a donc défini une nouvelle fonction économique — l'optimisation du rendement en sucre (pour un matériel donné), et construit un modèle mathématique tenant compte de la variation de la teneur en sucre de la betterave durant le ramassage et l'emmagasinage.

Une analyse ultérieure a révélé un défaut de ce modèle. Pour un horaire optimal (selon ce modèle) de travail, le ramassage devait commencer dès la période de croissance active de la plante, ce qui désavantageait économiquement les cultivateurs. Il fallait donc que le modèle prenne en considération l'intérêt de ces derniers. En résolvant un problème d'optimisation on doit viser par conséquent non seulement un plan de production, mais aussi un système de paye optimal.

Dans les problèmes mathématiques on s'interroge également sur l'objet de l'optimisation. Voici un exemple typique. Une fonction est donnée par une expression analytique compliquée ou par un tableau; on demande de l'approcher avec une précision donnée par une expression simple.

Etudions ce problème pour la fonction

$$f(q)=1-\frac{8}{\pi^2}\sum_{k=1}^{\infty}\frac{\exp\left(-\frac{2k-1}{2}\pi q\right)}{(2k-1)^2},\quad 0\leqslant q\leqslant\infty.$$

Effectuons pour simplifier le changement de variables

$$x=\exp\left(-\frac{\pi q}{2}\right),$$

i.e. considérons le problème d'approximation de la fonction

$$g(x)=1-\frac{8}{\pi^2}\sum_{k=1}^{\infty}\frac{x^{2k-1}}{(2k-1)^2}\quad \text{sur } [0,\ 1].$$

Les tentatives d'approcher par des polynômes ont été infructueuses, car, l'analyse l'a montré, $g'(x)$ n'était pas bornée au voisinage de $x=1$. L'étude de la singularité de la dérivée dans ce voisinage a suggéré de chercher l'approximation sous la forme

$$h(x)=(1-x)(a_0+a_1x+a_2x^2)\ln\frac{1+x}{1-x}+x(a_3+a_4x)+1.$$

Grâce à un choix heureux des paramètres a_i on a réussi à approcher la fonction à 10^{-4} près.

D'autres essais ont montré que pour $x\geqslant 10^{-4}$ $g(x)$ pouvait être approchée à $5\cdot10^{-4}$ près par une expression de la forme

$$\frac{(1-x)(1+x(a_1+a_2x))}{1+a_3x+a_4x^2+a_5x^3}.$$

Une fois qu'on a choisi la classe d'objets dans laquelle on cherche la solution du problème, il faut *paramétriser* convenablement cette classe. Ainsi, le polynôme

$$Q(x)=\sum_{i=0}^{n}a_ix^i,$$

qui approche $f(x)$ sur le segment $[-1,\ 1]$, s'écrit

$$Q(x)=\frac{b_0}{\sqrt{2}}+\sum_{i=1}^{n}b_iT_i(x).$$

La première écriture est, nous l'avons déjà dit, à bannir à cause d'une erreur de calcul éventuellement importante. La deuxième écriture présente un autre avantage. Supposons que le polynôme approximant est défini à partir de la condition

$$\min_Q\Phi,\quad \text{où}\quad \Phi=\int_{-1}^{1}\frac{(f(x)-Q(x))^2}{\sqrt{1-x^2}}dx.$$

En acceptant la première écriture, les surfaces de niveau de $\Phi(a_0, \ldots, a_n) = \text{const}$ sont des ellipsoïdes dont les axes sont très dispersés ; aussi les méthodes itératives de minimisation de cette fonction sont-elles lentement convergentes. Avec la deuxième écriture on a

$$\Phi(b_0, \ldots, b_n) = \int_{-1}^{1} \frac{(f(x))^2}{\sqrt{1-x^2}}\,dx - \sqrt{2}\,b_0 \int_{-1}^{1} \frac{f(x)}{\sqrt{1-x^2}}\,dx - \\ -2\sum_{i=1}^{n} b_i \int_{-1}^{1} \frac{f(x)\,T_i(x)}{\sqrt{1-x^2}}\,dx + \frac{\pi}{2}\sum_{i=0}^{n} b_i^2;$$

les surfaces de niveau de $\Phi(b_0, \ldots, b_n)$ sont des sphères et les procédés de minimisation par itérations garantissent une rapide convergence vers le point de minimum (ce point est connu en l'occurrence).

Un bon choix des paramètres et des critères de leur définition est particulièrement important lors du traitement des résultats d'observation.

Prenons pour modèle unidimensionnel du mouvement d'une avalanche de neige le système d'équations aux dérivées partielles

$$\frac{\partial h}{\partial t} + \frac{\partial hv}{\partial s} = 0,$$

$$\frac{\partial v}{\partial t} + v\frac{\partial v}{\partial s} + \frac{g}{2h}\frac{\partial h^2 \cos\psi}{\partial s} = g(\sin\psi - \mu\cos\psi) - k\frac{v^2}{h}; \qquad (1)$$

les conditions aux limites au front avant de l'avalanche s'écrivent

$$h(w-v) = h_0 w_0, \qquad h_0 w_0 v = \frac{1}{2} g h^2 \cos\psi - \sigma h,$$

où h est l'épaisseur de la neige, v la vitesse de l'avalanche, ψ la pente, w la vitesse du front avant et μ, k, σ des paramètres inconnus.

On a essayé de déterminer ces derniers paramètres en observant le front avant de l'avalanche qui se déplaçait sur un versant de pente constante $\psi = \text{const}$.

On a essayé de déterminer ces paramètres à partir de la condition de minimum de

$$\Phi(\mu, k, \sigma) = \sum_i (s(\mu, k, \sigma, t_i) - s_i)^2;$$

s_i désignent les positions observées du front avant aux instants t_i et $s(\mu, k, \sigma, t_i)$ les positions obtenues par intégration numérique du système (1). L'application directe d'un nombre important de méthodes de minimisation n'a révélé aucune tendance à la stabilisation de la part des valeurs calculées de μ, k, σ. Une analyse du problème a permis d'en découvrir la cause. C'est que pour $\psi = \text{const}$ l'égalité approchée $s(\mu, k, \sigma, t) \approx \lambda t$ est remplie avec un degré d'exactitude élevé ; dans cette égalité $\lambda = \left(1 + \frac{1}{\sqrt{\beta}}\right)^{3/2}$, $\beta = \frac{k}{\operatorname{tg}\psi - \mu}$. Aussi

la fonction $\Phi(\mu, k, \sigma)$ est très voisine d'une fonction $\Phi_0(\lambda)$. En minimisant Φ en μ, k, σ, on peut donc espérer obtenir non pas les valeurs de tous les paramètres μ, k, σ correspondant au point de minimum, mais seulement la valeur λ correspondant à ce point. Supposons que $\Phi_0(\lambda)$ est atteinte pour $\lambda = \lambda_0$. La surface $\lambda = \lambda_0$ est bidimensionnelle, et, dans notre cas, c'est une surface cylindrique dans l'espace des variables μ, k, σ. Sur cette surface la fonction $\Phi(\mu, k, \sigma)$ est approximativement constante et sa minimisation numérique détermine un déplacement en apparence chaotique des points (μ, k, σ) sur $\lambda = \lambda_0$. La véritable raison de notre échec n'est donc pas dans les défauts des procédés de minimisation, mais dans l'*insuffisance de l'information par rapport au but fixé*.

Mais passons du concret au général.

Quand on construit le modèle d'un problème, on a envie d'élaborer un modèle mathématique détaillé qui prenne en considération de nombreux éléments du problème, puis d'optimiser celui-ci complètement en choisissant au mieux tous les paramètres. Cette voie est riche en écueils.

Les premières difficultés apparaissent dès qu'on se met à décrire le modèle; en voulant prendre en considération trop de choses à la fois on risque de passer des faits importants et de nuire ainsi à la qualité du modèle. Lorsqu'on procède à une optimisation relativement à un grand nombre de paramètres, on a à minimiser une fonction de nombreuses variables. Or, en résolvant numériquement ce problème on se heurte parfois à des obstacles insurmontables.

Supposons que nous avons tout de même réussi à minimiser une telle fonction. Le problème suivant est de communiquer l'information recue au client afin de mettre en œuvre les résultats de l'analyse mathématique du modèle. En prenant la décision le client se guide généralement sur ses critères qualitatifs personnels sans s'intéresser aux détails du modèle. Si bien que si l'information qui lui parvient porte sur un nombre trop grand de paramètres, il n'est pas en mesure de se faire une opinion sur les recommandations et il y a de fortes chances qu'il les refuse purement et simplement.

Aussi, dans les premières phases d'optimisation, il importe surtout de construire un modèle élémentaire reflétant les paramètres principaux, déterminants. Cette tactique a également une autre raison d'être. Au début le client éprouve généralement de l'appréhension à l'égard des mathématiciens qui abordent en novices le domaine qui leur est proposé. Ce n'est qu'après avoir gagné sa confiance en obtenant de bons résultats dans l'optimisation des paramètres essentiels, qu'il vaut la peine de tenir compte de l'influence des paramètres négligés parmi les plus importants. Il ne faut pas non plus perdre de vue que la construction même d'un modèle amélioré du phénomène ne devient d'ordinaire possible qu'après que le client et l'utilisateur analysent en commun les résultats du calcul du modèle simplifié.

Il importe également que les résultats des recherches, lorsqu'on a à les discuter de concert avec des non-mathématiciens, soient

suffisamment suggestifs et présentés avec un appareil mathématique réduit à sa plus simple expression. Un recours injustifié à des notions mathématiques abstraites s'avère rarement d'une utilité quelconque.

§ 8. Comment optimiser ?

Une fois que le modèle et la fonction économique sont choisis et le problème paramétrisé, on doit minimiser une fonction (d'ordinaire de plusieurs variables) dans un domaine avec de nombreuses contraintes — égalités ou inégalités. La présence de contraintes complique singulièrement le problème de minimisation : il arrive généralement que le point d'extrémum est un point frontière du domaine. Vu l'extrême complexité du problème et l'existence d'une littérature spéciale abondante nous passerons outre aux divers raffinements relatifs à la minimisation dans un domaine avec contraintes et nous nous bornerons à des questions d'ordre général.

Les problèmes d'optimisation jouant un grand rôle dans les domaines d'activités les plus variés, on dispose actuellement d'un arsenal fourni de méthodes et de programmes standard correspondants. C'est pourquoi un problème concret isolé fait le plus souvent appel à l'un des programmes usuels disponibles.

Lorsqu'on aborde un problème nouveau, il est utile de prendre en considération les faits suivants. En ce qui concerne les méthodes examinées dans le présent chapitre la convergence des approximations a toujours été démontrée sous l'hypothèse d'une approximation initiale suffisamment bonne. Cette contrainte imposée à la méthode est intrinsèque ; ainsi, la surface de niveau d'un polynôme de degré n en m variables peut renfermer environ n^m composantes indépendantes dont la position relative est assez compliquée, et il est donc pratiquement impossible de construire un algorithme permettant de trouver rapidement le minimum d'un polynôme quelconque du huitième degré en dix variables.

On en tiendra compte en utilisant des programmes standard dont la description stipule la possibilité de minimiser des fonctions d'un *très* grand nombre de variables. *Dans ce meilleur des cas* il se peut que

a) le programme résolve efficacement des problèmes de minimisation appartenant à une classe que l'auteur du programme a l'habitude de manier ou que

b) formellement parlant, le programme soit apte à aborder tout problème de minimisation mais le temps de résolution est prohibitif dans la plupart des cas.

Lorsqu'on est confronté à un nouveau problème, on est souvent obligé d'élaborer des méthodes de recherche de l'approximation initiale qui sont spécifiques à la classe étudiée. L'exemple ci-dessous vous montrera que la recherche d'une approximation initiale acceptable est un problème ardu.

Il existe plusieurs programmes usuels de meilleure approximation de fonctions par le rapport

$$\frac{P_m(x)}{Q_n(x)} \tag{1}$$

de polynômes de degrés m et n donnés. En élaborant un programme standard de calcul des intégrales de Fresnel on s'est livré aux expériences suivantes. Chacun des programmes a été mis à l'essai en vue d'approcher la fonction considérée sur un segment pour tous les m et n compris dans les limites $0 < m + n \leqslant \leqslant 12$. Dans 3 cas au moins sur 90 chaque programme renvoyait à l'étiquette IMPOSSIBLE; dans certains programmes on avait plus de 50 % de refus. D'autre part, pour tous ces programmes, on a démontré des théorèmes de convergence sous l'hypothèse d'une approximation initiale assez bonne. On peut donc dire qu'il n'existe pas encore de programme standard, efficace et universel, de recherche de la meilleure approximation d'une fonction par des fractions rationnelles de la forme (1).

Afin de ne pas faire accroire que la résolution de problèmes d'optimisation tant soit peu compliqués est une entreprise désespérée, voyons la même question d'un œil plus optimiste.

Vus sous cet angle, nos arguments à l'appui de la complexité de la minimisation des polynômes paraissent peu convaincants. En effet, on ne nous demandera jamais de minimiser un polynôme *arbitraire*. Certains pensent que les problèmes de minimisation de fonctions possédant des courbes de niveau d'une structure *très* compliquée sont rares.

Considérons un exemple de problème où la nécessité même de sa résolution peut être mise en doute.

L'impossibilité d'obtenir une bonne approximation initiale peut s'expliquer par le comportement de la fonction considérée: le point de minimum se trouve dans un « puits » très étroit et en s'en éloignant dans toutes les directions la fonction s'accroît brusquement, puis commence à décroître (fig. 7.8.1). Supposons que ce minimum est réalisé en se déplaçant le long de grad Φ, en d'autres termes en intégrant le système

$$\frac{dx}{dt} + \operatorname{grad} \Phi = 0. \tag{2}$$

L'ensemble des conditions initiales à partir desquelles on arrive à ce point de minimum est contenu dans un petit voisinage de ce dernier. On peut dire qu'en adoptant pour le voisinage du point de minimum une échelle microscopique ce point s'avère un point attractif des solutions du système (2); pour une échelle plus grande, c'est un point répulsif (fig. 7.8.2). Ce comportement des approximations successives du point de minimum est également propre aux autres méthodes itératives.

Si le « puits » dont nous avons parlé est très étroit, il ne vaut parfois pas la peine de chercher le point de minimum. En effet, supposons que les paramètres x_i correspondent à un système de

commande réel ; le fonctionnement de ce système sera perturbé par certaines variations de ces paramètres. Si le point de minimum de la fonction Φ_i se trouve dans un tel « puits » étroit, de faibles variations des paramètres sont susceptibles de détériorer sensiblement

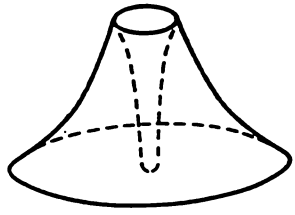


Fig. 7.8.1

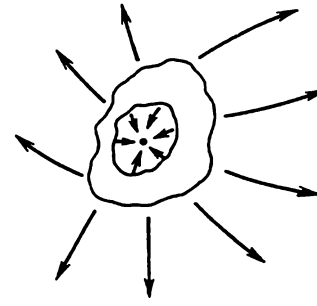


Fig. 7.8.2

les performances du système. A la lumière de ces faits, le choix du point d'extrémum dans le cas représenté par la figure 7.8.3 demande un examen complémentaire. Les chercheurs qui ont l'expérience

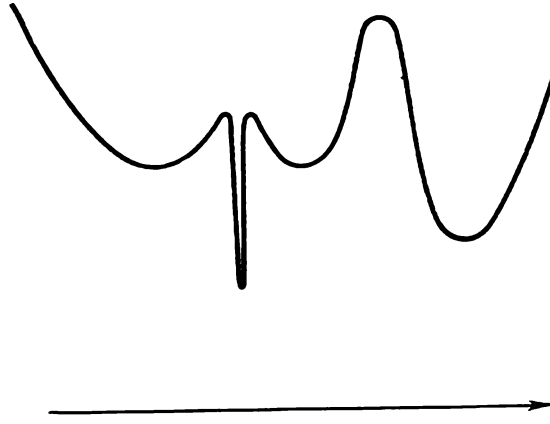


Fig. 7.8.3

des problèmes d'optimisation affirment que les fonctions économiques présentant des « puits » étroits sont le résultat d'une mauvaise construction du modèle mathématique.

Mais nous avons des raisons plus sérieuses d'être optimistes. Dans de nombreux cas la création d'un modèle mathématique et son optimisation ont pour but d'améliorer le fonctionnement d'un système existant. Les paramètres de celui-ci sont souvent une bonne approximation pour démarrer l'optimisation.

L'élaboration d'un nouveau système dicte d'ordinaire la tactique suivante. On commence par construire et optimiser un modèle très simple reflétant les facteurs essentiels. Le modèle se complique progressivement par adjonction de nouveaux facteurs. Il nous faut donc considérer de nouveaux problèmes de minimisation de fonctions dépendant d'un nombre toujours croissant de paramètres. Si les modèles auxiliaires sont construits convenablement, la solution de chaque problème d'optimisation constitue en général une bonne approximation initiale pour un autre problème plus compliqué.

Cette circonstance est souvent mise à profit de la manière suivante. Soit à minimiser une fonction dépendant d'un grand nombre de paramètres $x_1, \ldots, x_n$. Formons une fonction dépendant d'un nombre moindre de paramètres $X_1, \ldots, X_m$, $m < n$, qui approche la première, autrement dit construisons un modèle simplifié avec les paramètres essentiels $X_1, \ldots, X_m$. Optimisons le modèle et prenons l'approximation obtenue pour approximation première de l'optimisation ultérieure. Cette simplification du modèle et l'introduction de nouveaux paramètres sont parfois effectuées à plusieurs reprises.

L'optimisation de fonctions (modèles) dépendant d'un nombre moindre de paramètres est une opération plus facile : les courbes de niveau de la fonction à minimiser sont de structure plus simple et la recherche de chaque valeur de la fonction exige un volume de calcul moindre.

EXEMPLE. On cherche $\min\limits_{x_1, x_2, x_3} \Phi(x_1, x_2, x_3)$:

$$\Phi(x_1, x_2, x_3) = 1{,}1x_1^2 - 2x_1x_2 + x_2^2 + x_3^4 + \exp(x_1x_3).$$

La fonction $\Phi_1(x_1, x_2, x_3)$ est proche de

$$\Phi_1(x_1, x_2, x_3) = (x_1 - x_2)^2 + x_3^4 + \exp(x_1x_3).$$

Evidemment

$$\min_{x_2} \Phi_1(x_1, x_2, x_3) = \Phi_2(x_1, x_3) = x_3^4 + \exp(x_1x_3).$$

La recherche de l'approximation initiale s'est ramenée à minimiser une fonction de deux variables.

La construction d'un modèle simplifié n'implique pas obligatoirement la diminution du nombre de paramètres.

Qu'entend-on par facteurs ou paramètres principaux? On peut dire que ce sont ceux dont la fonction dépend fortement; en termes des mathématiques, cela signifie que les dérivées de la fonction par rapport à ces paramètres sont relativement grandes. Les paramètres secondaires sont ceux dont la fonction en question dépend faiblement, i. e. tels que les dérivées par rapport à ces paramètres sont petites. Si la dérivée de $\Phi(x_1, \ldots, x_n)$ par rapport à x_n est petite, cela veut dire que celle-ci est proche de la fonction $\Phi(x_1, \ldots, x_{n-1}, \bar{x}_n)$, $\bar{x}_n = \text{const}$, qui dépend d'un nombre moindre de

paramètres. On définit formellement l'importance des paramètres en appréciant les dérivées de la fonction. Cette évaluation est cependant laborieuse pour des problèmes compliqués; une caractérisation des paramètres selon le critère plus concret qu'est leur importance permet au responsable ou au chercheur de suggérer au mathématicien ceux qu'il faut choisir en premier lieu.

On aboutit parfois à une méthode d'optimisation acceptable ou à un procédé de recherche d'une bonne approximation initiale en étudiant les principes dont s'inspirent un praticien ou un dirigeant expérimentés, ou bien leurs méthodes de travail.

L'exemple suivant prouve combien il est utile de profiter de l'expérience des praticiens.

Une usine qui venait à peine d'entrer en service souffrait de mauvaises cadences de production par la faute d'opérateurs peu expérimentés. On a essayé en vain d'élaborer un modèle de processus de production qui soit suffisamment exact et se prête à une analyse mathématique sur calculateurs. On a donc abandonné momentanément l'espoir de créer un modèle et accepté un autre moyen d'automatiser et d'optimiser la production. On a stocké en mémoire de la machine les régimes de travail des meilleurs opérateurs des entreprises connexes. Puis, selon les conditions présentes, la machine choisissait le régime de travail le plus proche de celui d'un des meilleurs opérateurs. Cette mesure a permis de lever les difficultés mentionnées.

Les procédés de recherche de l'approximation initiale et les méthodes itératives mêmes présentent souvent une analogie avec un processus réel dans le cadre d'un phénomène différent.

Nous avons évoqué plus haut la méthode de la boule pesante et d'autres procédés de stationnarisation. L'activité du cerveau et de l'appareil optique en quête d'un objet quelconque suit probablement le schéma suivant. On procède d'abord à un survol du champ visuel à une grande échelle, puis on sélectionne à l'aide de l'information reçue le secteur concerné à la même échelle et ainsi de suite. Cette façon d'agir fait penser aux méthodes des §§ 17, 18 du chap. III et à celles du cinquième paragraphe du présent chapitre.

Lorsque la décision est prise de poursuivre l'étude de tels problèmes nouveaux ardus, l'organisation de la recherche est d'une grande importance. Si l'on sait dans quelle direction prospecter, l'on concentre tous ses efforts en conséquence. Si, par contre, on hésite sur la voie à prendre, il faut alors doubler les recherches, i. e. confier le problème à plusieurs organisations qui chercheront chacune de leur côté et à leur manière, sans parfois échanger systématiquement d'information. Car s'il semble de prime abord que l'échange d'information est utile, un échange constant risque par contre d'empêcher la gestation et le développement de toute solution originale.

Selon une hypothèse (qui n'est pas universellement admise), c'est également la façon dont fonctionne le cerveau quand il a à résoudre un problème: l'information obtenue se fixe avec une certaine indétermination dans différentes zones du cerveau; d'autre

part, à chaque instant, cette information est fournie par un secteur localisé du cerveau en activité.

Par analogie avec ce que nous avons dit plus haut, si l'on a un problème d'optimisation (et en général un problème quelconque) à résoudre d'urgence, on procédera comme suit. On essaie successivement ou simultanément plusieurs méthodes de résolution de problèmes analogues. Qu'ils soient utilisés à tour de rôle ou tous à la fois, les algorithmes A_p, $p = 1, \ldots, l$, travaillent par cycles (l'intervalle de temps t_{pq} durant lequel l'algorithme A_p s'applique la q-ième fois est donné par l'utilisateur ou se définit en cours de travail). Dans une utilisation à tour de rôle chaque algorithme démarre à partir de l'approximation obtenue par son prédécesseur; dans une utilisation simultanée chaque algorithme commence la minimisation à partir de la valeur approchée trouvée par ce même algorithme au bout du cycle précédent (on dirait dans ce cas que les algorithmes travaillent « à qui mieux mieux »).

L'utilisation indépendante des algorithmes est l'analogie d'un travail en parallèle sans échange d'information. Quant à l'analogue de la recherche avec échange d'information aux instants discrets, cela peut être le régime de travail suivant. Au début de chaque intervalle t_{pq} l'algorithme A_p examine un certain lot d'approximations réalisées par tous les algorithmes et en choisit celle qui lui paraît être la meilleure. Ce lot peut être formé de toutes les approximations du cycle précédent ou d'approximations obtenues aux extrémités de tous les t_{ij}.

Le régime suivant peut s'avérer rentable : une fois que A_p a obtenu une approximation qui est très bonne pour A_s, on alloue le temps à ce dernier. On ne doit certes pas penser qu'une imitation pure et simple de systèmes réels permet toujours de bien résoudre le problème d'optimisation proposé.

Faisons un bref bilan. Les problèmes d'optimisation des fonctions d'un grand nombre de variables sont généralement très ardus; dans le cas d'un problème nouveau on se dépense, et parfois en vain, à tester des algorithmes connus ou nouveaux. Pourtant si l'on garde un contact permanent avec les praticiens et si l'on a la possibilité d'étudier en détail les modèles simplifiés, on peut être sûr d'aboutir.

A l'heure actuelle on note un progrès sensible dans la résolution des problèmes d'optimisation grâce à l'emploi des systèmes conversationnels (où l'homme, mathématicien ou responsable, dialogue avec la machine) fonctionnant en régime partagé.

Bibliographie du chapitre VII

II [10, 21, 25, 42, 47]; § 2: III [29]; §§ 3, 4: I [2], II [21, 32]; § 5: II [32] *), III [19]; §§ 7, 8: II [21, 40, 41, 43], III [47].

*) Voir également F. Vassiliev, « Conférences sur les méthodes de résolution des problèmes intégraux », éd. Université de Moscou, 1974 (éd. russe).

TROISIÈME PARTIE

MÉTHODES DE RÉSOLUTION NUMÉRIQUE DES ÉQUATIONS DIFFÉRENTIELLES ORDINAIRES

CHAPITRE VIII

MÉTHODES DE RÉSOLUTION NUMÉRIQUE DU PROBLÈME DE CAUCHY

La résolution des équations différentielles ordinaires est plus difficile que le calcul d'intégrales simples, et la part des problèmes intégrés explicitement est ici sensiblement moindre.

En parlant de l'intégrabilité sous forme explicite on sous-entend généralement que la solution s'obtient au bout d'un nombre fini d'opérations « élémentaires »: addition, soustraction, multiplication, division, exponentiation, logarithmes et opération inverse, calcul du sinus et du cosinus. La notion d'opération et de fonction « élémentaires » a changé dès avant l'avènement des machines à calculer électroniques. On a eu si souvent affaire dans les applications aux solutions de certains problèmes particuliers qu'on a formé les tableaux de leurs valeurs, par exemple les tableaux des intégrales de Fresnel

$$\int_0^x \exp(it^2)\,dt,$$

du sinus intégral $\int_0^x \frac{\sin t}{t}\,dt$, des solutions de l'équation différentielle

$$x^2y'' + xy' + (x^2 - \nu^2)\,y = 0,$$

appelées fonctions de Bessel, et de plusieurs autres fonctions. Ces fonctions sont dites spéciales. Avec de tels tableaux, en principe il n'y a plus aucune différence entre la recherche de $\sin x$, $\ln x$, ... et celle des fonctions spéciales. Les unes et les autres se calculent à l'aide des tableaux, sont approximées par des polynômes, des fonctions rationnelles, etc. Ainsi, dans la classe de problèmes intégrables explicitement sont entrés des problèmes dont les solutions s'expriment par des fonctions spéciales. Cependant, même ainsi élargie, cette classe ne représente qu'une partie relativement faible de problèmes proposés. L'élaboration des méthodes numériques et la généralisation des ordinateurs ont déterminé une large extension de la classe d'équations différentielles réelles et donc de la sphère d'application des mathématiques.

A l'heure actuelle, le travail préparatoire nécessaire pour résoudre sur machine une équation différentielle n'est pas tellement supérieur à l'effort de mise en équation même. On obtient à volonté à la sortie de la machine soit un graphique imprimé de la solution, soit une courbe représentative visualisée sur un écran. On n'a plus besoin d'étudier de nombreux procédés particuliers d'intégration sous forme explicite.

Le présent chapitre traite de diverses questions relatives à la théorie des méthodes numériques pour les équations différentielles, méthodes qui ont été en particulier mises à la base des programmes standard de résolution des équations différentielles ordinaires.

§ 1. Développement de la solution en série de Taylor

La méthode par la série de Taylor est l'une des plus anciennes techniques de résolution d'équations différentielles.

On désire trouver sur le segment $[x_0, x_0 + X]$ la solution de

$$y' = f(x, y) \tag{1}$$

avec la condition initiale $y(x_0) = y_0$; $f(x, y)$ est analytique au point (x_0, y_0). La dérivation de (1) par rapport à x fournit

$$y'' = f_x(x, y) + f_y(x, y)\, y',$$
$$y''' = f_{xx}(x, y) + 2f_{xy}(x, y)\, y' + f_{yy}(x, y)\, y'^2 + f_y(x, y)\, y'', \ldots$$

En substituant $x = x_0$ et $y = y_0$, on obtient les valeurs consécutives

$$y'(x_0),\ y''(x_0),\ y'''(x_0),\ \ldots$$

On peut donc écrire l'égalité approchée

$$y(x) \approx \sum_{i=0}^{n} \frac{y^{(i)}(x_0)}{i!}(x - x_0)^i. \tag{2}$$

Si $|x - x_0|$ est plus grand que le rayon de convergence de la série

$$\sum_i \frac{y^{(i)}(x_0)}{i!}(x - x_0)^i,$$

l'erreur (2) ne tend pas vers zéro lorsque $n \to \infty$. Un procédé plus rentable consiste à partager $[x_0, x_0 + X]$ en segments partiels

$$[x_{j-1}, x_j], \quad j = 1, \ldots, N;$$

on trouve de proche en proche les approximations y_j de la solution $y(x_j)$, $j = 1, \ldots, N$, d'après la règle suivante: supposons y_j déjà calculé; calculons les valeurs que prennent en x_j les dérivées $y_j^{(i)}$ de la solution de l'équation différentielle initiale, qui passe par (x_j, y_j); posons

$$y(x) \approx z_j(x) \approx \sum_{i=0}^{n} \frac{y_j^{(i)}}{i!}(x - x_j)^i \tag{3}$$

sur le segment $[x_j, x_{j+1}]$ et, partant,

$$y_{j+1} = z_j(x_{j+1}). \tag{4}$$

Plus loin nous discuterons en détail l'évaluation de l'erreur dans les méthodes d'intégration à pas séparés dont fait partie le procédé considéré. Pour l'instant nous nous contenterons d'expliquer pourquoi celui-ci peut donner lieu à une erreur faible. Considérons le cas $x_{j+1} - x_j \equiv h$; si y_j coïncidait avec la valeur exacte $y(x_j)$, l'erreur commise en remplaçant y_{j+1} par $z_j(x_{j+1})$ serait en $O(h^{n+1})$. Puisque l'erreur est introduite sur $O(h^{-1})$ segments, le fractionnement du pas du réseau entraînera visiblement

$$\max_{0 \leqslant j \leqslant N} |y_j - y(x_j)| = O(h^n).$$

Nous avons là un exemple typique de raisonnements vraisemblables visant à légitimer telle ou telle méthode numérique; or, il arrive souvent que la convergence de la méthode est prouvée par de pareils raisonnements alors que la solution approchée ne converge pas en réalité vers la solution exacte. Aussi est-il important, sur le plan théorique comme sur le plan pratique, de justifier rigoureusement la convergence d'une méthode avec fractionnement du pas et d'évaluer l'erreur.

Voilà déjà plus de cent ans qu'on a abandonné le procédé décrit d'intégration des équations différentielles pour $n > 1$. En effet, dans ce procédé on calcule les valeurs de la fonction f et de toutes ses dérivées $f_x{}^j{}_y{}^{m-j}$ pour $m < n$, i.e. on trouve $\frac{n(n+1)}{2}$ fonctions différentes. A part certaines autres conséquences fâcheuses, cela a pour effet, en calcul manuel, de disperser l'attention et d'augmenter la probabilité d'erreur. En calcul automatique, la méthode exige l'écriture de nombreux blocs de calcul des dérivées, ce qui contredit la tendance fondamentale à simplifier les relations entre la machine et les utilisateurs.

Il existe des systèmes de programmes qui réalisent une dérivation sur formules; pour un programme donné de calcul de $f(x, y)$ ce système détaille les programmes de calcul de ses dérivées et on n'a donc pas besoin de les rédiger. En ce qui concerne le coût en temps machine, ces algorithmes le cèdent beaucoup aux méthodes de Runge-Kutta qui seront examinées plus bas et aux méthodes des différences finies.

Nous ne voulons cependant pas que le lecteur prenne nos raisonnements pour une invitation expresse à proscrire les méthodes utilisant le développement taylorien. Dans certains cas on résout sur ordinateur des équations différentielles d'un type strictement déterminé. Le calcul des trajectoires des corps célestes, par exemple, se fait par une intégration multiple de systèmes d'équations différentielles d'une forme déterminée avec diverses conditions initiales et les paramètres différents des seconds membres. Le fait d'avoir affaire à un même système présente souvent un certain avantage, à savoir les formules concrètes donnant les dérivées des seconds membres ont d'ordinaire de nombreux termes en commun, et il peut arriver, pour telle ou telle classe de systèmes, que le calcul simultané de toutes ces dérivées demande des opérations arithmétiques en nombre relativement faible et que le mode opératoire considéré soit plus efficace que d'autres procédés d'intégration numérique. Dans la suite, nous nous bornerons essentiellement à des méthodes qui nécessitent uniquement le calcul des valeurs du second membre; ces méthodes sont généralement à la base des programmes standard de résolution de vastes classes de problèmes types.

§ 2. Méthodes de Runge-Kutta

Dans le cas particulier $n = 1$, la formule (1.3) s'écrit

$$y_{j+1} = y_j + hf(x_j, y_j), \quad h = x_{j+1} - x_j. \tag{1}$$

C'est la *méthode d'Euler*. On peut construire une autre classe de formules de calcul renfermant cette méthode. Citons d'abord les procédés les plus simples de cette classe qu'on obtient intuitivement. Supposons que l'on connaisse une valeur de la solution $y(x)$ et qu'on veuille trouver la valeur $y(x + h)$. Considérons l'égalité

$$y(x+h) = y(x) + \int_0^h y'(x+t)\,dt. \tag{2}$$

En remplaçant l'intégrale à droite par $hy'(x)$, on commet une erreur en $O(h^2)$, i.e.

$$y(x + h) = y(x) + hy'(x) + O(h^2);$$

vu que $y'(x) = f(x, y(x))$, il en découle

$$y(x + h) = y(x) + hf(x, y(x)) + O(h^2).$$

On aboutit à la formule d'Euler (1) en rejetant le terme en $O(h^2)$ et en notant $x = x_j$, $x + h = x_{j+1}$. On obtient une formule plus exacte en approchant mieux l'intégrale dans le second membre de (2). La formule des trapèzes donne

$$y(x+h) = y(x) + \frac{h}{2}(y'(x) + y'(x+h)) + O(h^3),$$

ou encore

$$y(x+h) = y(x) + \frac{h}{2}(f(x, y(x)) + f(x+h, y(x+h))) + O(h^3). \tag{3}$$

La formule de calcul correspondante a la forme :

$$y_{j+1} = y_j + \frac{h}{2}(f(x_j, y_j) + f(x_{j+1}, y_{j+1})). \tag{4}$$

En général, cette équation n'admet pas de solution explicite par rapport à y_{j+1}. Continuons donc de transformer l'algorithme. La substitution de

$$y^* = y(x + h) + O(h^2) \tag{5}$$

à $y(x + h)$ à droite dans (3) fait que le second membre varie de

$$\frac{h}{2}(f(x+h, y^*) - f(x+h, y(x+h))) = \frac{h}{2} f_y(x+h, \bar{y})(y^* - y(x+h)),$$

avec $\bar{y}$ compris entre y^* et $y(x+h)$; sous l'hypothèse (5) cette quantité est de l'ordre de $O(h^3)$. On a donc

$$y(x+h) = y(x) + \frac{h}{2}(f(x, y(x)) + f(x+h, y^*)) + O(h^3);$$

la condition (5) vérifie le résultat des calculs selon la formule d'Euler :

$$y^* = y(x) + hf(x, y(x)).$$

Les dernières relations déterminent un couple de formules de calcul

$$\begin{aligned} y^*_{j+1} &= y_j + hf(x_j, y_j), \\ y_{j+1} &= y_j + \frac{h}{2}(f(x_j, y_j) + f(x_{j+1}, y^*_{j+1})). \end{aligned} \tag{6}$$

Construisons un autre couple de formules avec une erreur par pas du même ordre de grandeur. Remplaçons l'intégrale dans le second membre de (2) selon la formule des rectangles

$$y(x+h) = y(x) + hy'\left(x + \frac{h}{2}\right) + O(h^3)$$

ou, ce qui revient au même,

$$y(x+h) = y(x) + hf\left(x + \frac{h}{2}, y\left(x + \frac{h}{2}\right)\right) + O(h^3).$$

Si

$$y^* = y\left(x + \frac{h}{2}\right) + O(h^2),$$

on a, comme dans le cas précédent,

$$y(x+h) = y(x) + hf\left(x + \frac{h}{2}, y^*\right) + O(h^3).$$

On prend en qualité de y^* le résultat obtenu à l'aide de la formule d'Euler avec le pas $\frac{h}{2}$: $y^* = y(x) + \frac{h}{2} f(x, y(x))$. A ces relations correspond le couple de formules

$$\begin{aligned} y_{j+\frac{1}{2}} &= y_j + \frac{h}{2} f(x_j, y_j), \\ y_{j+1} &= y_j + hf\left(x_j + \frac{h}{2}, y_{j+\frac{1}{2}}\right). \end{aligned} \tag{7}$$

Les procédés obtenus se rapportent aux méthodes de Runge-Kutta du type suivant. En cours de calcul on fixe certains nombres

$$\alpha_2, \ldots, \alpha_q, \quad p_1, \ldots, p_q, \quad \beta_{ij}, \ 0 < j < i \leqslant q;$$

on trouve successivement

$$k_1(h) = hf(x, y),$$
$$k_2(h) = hf(x + \alpha_2 h, y + \beta_{21} k_1(h)),$$
$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$$
$$k_q(h) = hf(x + \alpha_q h, y + \beta_{q,1} k_1(h) + \ldots + \beta_{q,q-1} k_{q-1}(h))$$

et on pose

$$y(x+h) \approx z(h) = y(x) + \sum_{i=1}^{q} p_i k_i(h).$$

Passons au choix des paramètres α_i, p_i, β_{ij}. Notons $\varphi(h) = y(x+h) - z(h)$. Si $f(x, y)$ est une fonction suffisamment régulière de ses arguments, alors $k_1(h), \ldots, k_q(h)$ et $\varphi(h)$ sont des fonctions régulières du paramètre h. Supposons que

$$\varphi(0) = \varphi'(0) = \ldots = \varphi^{(s)}(0) = 0$$

pour $f(x, y)$ arbitraires suffisamment régulières et que $\varphi^{(s+1)}(0) \neq 0$ pour une certaine fonction régulière $f(x, y)$. Selon la formule de Taylor,

$$\varphi(h) = \sum_{i=0}^{s} \frac{\varphi^{(i)}(0)}{i!} h^i + \frac{\varphi^{(s+1)}(\theta h)}{(s+1)!} h^{s+1} = \frac{\varphi^{(s+1)}(\theta h)}{(s+1)!} h^{s+1} \qquad (8)$$

où $0 < \theta < 1$. La quantité $\varphi(h)$ est appelée erreur de méthode sur un pas et s son ordre de grandeur. Lorsque $q = 1$, on a

$$\varphi(h) = y(x+h) - y - p_1 hf(x, y),$$
$$\varphi(0) = 0, \quad \varphi'(0) = (y'(x+h) - p_1 f(x, y))|_{h=0} =$$
$$= f(x, y)(1 - p_1), \quad \varphi''(h) = y''(x+h);$$

ici et dans la suite $y = y(x)$. L'égalité $\varphi'(0) = 0$ est évidemment vraie pour toutes les $f(x, y)$ dans le seul cas $p_1 = 1$. La méthode d'Euler correspond à cette dernière valeur. Conformément à (8), l'erreur par pas dans cette méthode s'écrit

$$\varphi(h) = \frac{y''(x+\theta h) h^2}{2}.$$

Soit $q = 2$; alors

$$\varphi(h) = y(x+h) - y - p_1 hf(x, y) - p_2 hf(\bar{x}, \bar{y})$$

où $\bar{x} = x + \alpha_2 h$, $\bar{y} = y + \beta_{21} hf(x, y)$.

Calculons les dérivées de $\varphi(h)$:

$$\varphi'(h) = y'(x+h) - p_1 f(x, y) - p_2 f(\bar{x}, \bar{y}) -$$
$$- p_2 h(\alpha_2 f_x(\bar{x}, \bar{y}) + \beta_{21} f_y(\bar{x}, \bar{y}) f(x, y)),$$

$$\varphi''(h) = y''(x+h) - 2p_2(\alpha_2 f_x(\bar{x}, \bar{y}) +$$
$$+ \beta_{21} f_y(\bar{x}, \bar{y}) f(x, y)) - p_2 h(\alpha_2^2 f_{xx}(\bar{x}, \bar{y}) +$$
$$+ 2\alpha_2\beta_{21} f_{xy}(\bar{x}, \bar{y}) f(x, y) + \beta_{21}^2 f_{yy}(\bar{x}, \bar{y})(f(x, y))^2),$$
$$\varphi'''(h) = y'''(x+h) - 3p_2(\alpha_2^2 f_{xx}(\bar{x}, \bar{y}) +$$
$$+ 2\alpha_2\beta_{21} f_{xy}(\bar{x}, \bar{y}) f(x, y) + \beta_{21}^2 f_{yy}(\bar{x}, \bar{y})(f(x, y))^2) + O(h).$$

En conformité avec l'équation différentielle initiale,

$$y' = f, \quad y'' = f_x + f_y f, \quad y''' = f_{xx} + 2f_{xy} f + f_{yy} f^2 + f_y y''.$$

Portons $h = 0$ dans les expressions de $\varphi(h)$, $\varphi'(h)$, $\varphi''(h)$, $\varphi'''(h)$ et utilisons les relations précédentes; il vient

$$\begin{aligned}
&\varphi(0) = y - y = 0,\\
&\varphi'(0) = (1 - p_1 - p_2) f(x, y),\\
&\varphi''(0) = (1 - 2p_2\alpha_2) f_x(x, y) + (1 - 2p_2\beta_{21}) f_y(x, y) f(x, y),\\
&\varphi'''(0) = (1 - 3p_2\alpha_2^2) f_{xx}(x, y) + (2 - 6p_2\alpha_2\beta_{21}) f_{xy}(x, y) f(x, y) +\\
&+ (1 - 3p_2\beta_{21}^2) f_{yy}(x, y)(f(x, y))^2 + f_y(x, y) y''(x).
\end{aligned} \tag{9}$$

$\varphi'(0) = 0$ quel que soit f si

$$1 - p_1 - p_2 = 0, \tag{10}$$

$\varphi''(0) = 0$ si

$$1 - 2p_2\alpha_2 = 0 \quad \text{et} \quad 1 - 2p_2\beta_{21} = 0. \tag{11}$$

Ainsi, $\varphi(0) = \varphi'(0) = \varphi''(0) = 0$ pour n'importe quelle $f(x, y)$ si sont réalisées les trois conditions ((10)-(11)). En se donnant de façon arbitraire l'un des quatres paramètres, on aboutit à une méthode de Runge-Kutta avec une erreur d'ordre infinitésimal 2 en h. Par exemple, lorsque $p_1 = 1/2$, on obtient $p_2 = 1/2$, $\alpha_2 = 1$, $\beta_{21} = 1$, ce qui correspond au couple (6). Lorsque $p_1 = 0$, on a $p_2 = 1$, $\alpha_2 = 1/2$, $\beta_{21} = 1/2$, ce qui correspond à (7). Dans le cas $y' = y$, il vient, en conformité avec (9), $\varphi'''(0) = y$ quels que soient p_1, p_2, α_2, β_{21}. Il en résulte qu'il n'existe pas de formules de Runge-Kutta avec $q = 2$ et $s = 3$.

Vu (8), la partie principale de l'erreur par pas dans les méthodes où $q = s = 2$ est $\frac{\varphi'''(0)}{6} h^3$. En se servant de l'expression explicite de $\varphi'''(0)$ on arrive souvent à réduire cette quantité par un choix convenable des paramètres. Si $f_y y''$ est faible pour la classe considérée d'équations, on choisira les paramètres de manière que les trois premiers termes de la dernière formule (9) s'annulent. On vérifie aisément que lorsque $p_1 = 1/4$, $p_2 = 3/4$, $\alpha_2 = \beta_{21} = 2/3$, on

a (10) et (11) et $\varphi'''(0) = f_y(x, y)\, y'(x)$. A cet ensemble de paramètres correspondent les formules de calcul

$$\begin{aligned} k_1 &= hf(x, y), \\ k_2 &= hf\left(x+\frac{2}{3}h,\ y+\frac{2}{3}k_1\right), \\ \Delta y &= z(h) - y = \frac{1}{4}(k_1+3k_2). \end{aligned} \tag{12}$$

Des raisonnements analogues pour $q = 3$ aboutissent aux résultats suivants. Comme il n'existe pas de formules de calcul correspondant à $s = 4$, pour que s vaille 3, il faut que

$$\alpha_2 = \beta_{21}, \quad \alpha_3 = \beta_{31}+\beta_{32}, \quad \alpha_3(\alpha_3-\alpha_2) - \beta_{32}\alpha_2(2-3\alpha_2) = 0,$$

$$p_3\beta_{32}\alpha_2 = \frac{1}{6}, \quad p_2\alpha_2 + p_3\alpha_3 = \frac{1}{2}, \quad p_1+p_2+p_3 = 1.$$

Ce système de six équations à huit inconnues admet une infinité de solutions. Le jeu de formules de calcul le plus usité est

$$k_1 = hf(x, y), \quad k_2 = hf\left(x+\frac{h}{2},\ y+\frac{k_1}{2}\right),$$

$$k_3 = hf(x+h,\ y-k_1+2k_2), \quad \Delta y = \frac{1}{6}(k_1+4k_2+k_3).$$

Si $f(x, y)$ ne dépend pas de y, i.e. $f_y \equiv 0$, on retrouve la formule de Simpson

$$y(x+h) - y(x) \approx \frac{h}{6}\left(f(x)+4f\left(x+\frac{h}{2}\right)+f(x+h)\right). \tag{13}$$

Selon l'estimation de l'erreur dans la formule de Simpson, cette égalité approchée est entachée d'une erreur en $O(h^5)$. Aussi, de même que (12), cette formule est surtout indiquée pour f_y petits.

Si $q = 4, 5$, on ne réussit pas à construire des formules avec $s = 5$; si $q = 4$, on a un ensemble biparamétrique de formules correspondant à $s = 4$. Citons à titre d'exemple une famille à un paramètre :

$$\begin{aligned} k_1 &= hf(x, y), \quad k_2 = hf\left(x+\frac{h}{2},\ y+\frac{k_1}{2}\right), \\ k_3 &= hf\left(x+\frac{h}{2},\ y+\left(\frac{1}{2}-\frac{1}{2t}\right)k_1+\frac{1}{2t}k_2\right), \\ k_4 &= hf(x+h,\ y+(1-t)k_2+tk_3), \\ \Delta y &= \frac{1}{6}(k_1+(4-2t)k_2+2tk_3+k_4). \end{aligned}$$

L'ensemble de formules le plus usité de cette famille correspond à la valeur $t = 1$:

$$k_1 = hf(x, y), \quad k_2 = hf\left(x+\frac{h}{2},\ y+\frac{k_1}{2}\right),$$

$$k_3 = hf\left(x+\frac{h}{2},\ y+\frac{k_2}{2}\right),\quad k_4 = hf(x+h,\ y+k_3), \tag{14}$$
$$\Delta y = \frac{1}{6}(k_1+2k_2+2k_3+k_4).$$

Indiquons un autre jeu de formules avec $q=s=4$:

$$\begin{aligned} k_1 &= hf(x,\ y),\quad k_2 = hf\left(x+\frac{h}{3},\ y+\frac{k_1}{3}\right),\\ k_3 &= hf\left(x+\frac{2h}{3},\ y-\frac{k_1}{3}+k_2\right),\\ k_4 &= hf(x+h,\ y+k_1-k_2+k_3),\\ \Delta y &= \frac{1}{8}(k_1+3k_2+3k_3+k_4). \end{aligned} \tag{15}$$

En disant d'un ensemble de formules qu'il est « le plus usité », nous avons suivi une tradition ancienne en analyse numérique. Il nous semble cependant qu'au lieu de refléter une tendance, un manuel de calcul numérique devrait plutôt indiquer la meilleure formule d'une famille donnée. Or, ce n'est pas si simple que cela.

Prenons pour commencer le problème de calcul d'une intégrale qui se prête le mieux à une comparaison des modes opératoires. Pour un même ordre de grandeur, les erreurs sur les formules de quadrature ont toujours leurs parties principales proportionnelles. Ainsi, en calculant $\int_0^1 f(x)\,dx$ avec partage de [0, 1] en segments partiels de longueur h on commet une erreur

$$\frac{(f'(1)-f'(0))\,h^2}{12}+O(h^3)$$

avec la formule des trapèzes, et

$$-\frac{(f'(1)-f'(0))\,h^2}{24}+O(h^3)$$

avec celle des rectangles; le nombre de points en lesquels on calcule $f(x)$ est sensiblement identique, soit $1+h^{-1}$ et h^{-1} respectivement. La comparaison des parties principales des erreurs incite à donner la préférence à la formule des rectangles. Cette conclusion n'est cependant pas définitive. En effet, l'estimation du reste comprend un terme infiniment petit d'ordre supérieur, et l'information dont nous disposons ne nous autorise à préférer la formule des rectangles que pour h très faibles. On a comparé plus haut ces méthodes en les mettant en œuvre une seule fois. Or, pour trouver une intégrale, il arrive souvent qu'on change de pas. Les points de base de la formule des trapèzes avec le pas $h' = hq^{-1}$, q entier, le restent pour un pas h, et l'intégration avec le pas hq^{-1} utilisera donc les résultats obtenus avec le pas h. Quant à la formule des rectangles, elle ne le permet que pour le pas $h' = hq^{-1}$, q impair. Vu que la valeur $q=2$ présente certaines commodités, l'avantage de la formule des rectangles n'est déjà plus si évident.

L'étude des équations différentielles fait apparaître plusieurs circonstances supplémentaires. Dans les formules exactes à un même ordre relativement à h,

les parties principales de l'erreur sur un pas s'avèrent souvent non proportionnelles. Ainsi, par suite de (8), (9), la partie principale de l'erreur sur (6) vaut $(B - A)\,h^3$, avec $B = \frac{1}{6} f_y y''$,

$$A = \frac{1}{12}\,(f_{xx} + 2f_{xy}y' + f_{yy}\,(y')^2),$$

et elle est $\left(B + \frac{A}{2}\right) h^3$ pour la formule (7). Par conséquent, on choisira la méthode ((6) ou (7)) selon l'équation proposée.

Considérons un exemple. On demande d'intégrer $y' = y^q$, $q > 0$, lorsque $y\,(0) > 1$. Les parties principales des erreurs dues à (6) et (7) sont respectivement

$$\frac{q\,(q+1)\,y^{3q-2}h^3}{12} \quad \text{et} \quad \frac{q\,(3q+1)\,y^{3q-2}h^3}{24};$$

si $q > 1$, la deuxième est supérieure à la première et vice versa si $q < 1$.

C'est le libre arbitre qui décide alors de la méthode, compte tenu des traditions et de la statistique des propriétés des problèmes proposés.

Voici un exemple de caractérisation comparative des méthodes. Soit à résoudre des équations différentielles aléatoires d'un certain ensemble. Supposons que

$$\xi_1 = f_{xx} + 2f_{xy}y' + f_{yy}\,(y')^2 \text{ et } \xi_2 = f_y y''$$

sont des variables aléatoires indépendantes telles que $E\xi_1 = E\xi_2 = 0$. L'espérance mathématique du carré de la partie principale de l'erreur par pas est alors respectivement

$$\left(\frac{1}{144} E\xi_1^2 + \frac{1}{36} E\xi_2^2\right) h^6, \qquad \left(\frac{1}{576} E\xi_1^2 + \frac{1}{36} E\xi_2^2\right) h^6.$$

Ici la formule (7) l'emporte sur (6). Mais la meilleure est la formule (12) où l'espérance mathématique correspondante vaut $\frac{1}{36} E\xi_2^2 h^6$. Nous avons raisonné dans l'hypothèse de ξ_1 et ξ_2 aléatoires et indépendants, une hypothèse, à vrai dire, douteuse. On doit donc faire preuve d'une grande prudence en discutant les conclusions obtenues.

Reprenons les formules (14) et (15). Dans le cas de $y' = f(x)$ l'erreur résultant de (14) s'écrit

$$\rho_1 = y\,(x+h) - y\,(x) - \frac{h}{6}\left(f\,(x) + 4f\left(x + \frac{h}{2}\right) + f\,(x+h)\right).$$

Le développement en série de Taylor donne

$$\rho_1 = -\frac{f^{(4)}\,(x)\,h^5}{2880} + O\,(h^6).$$

Et pour la formule (15)

$$\rho_2 = y\,(x+h) - y\,(x) - \frac{h}{8}\left(f\,(x) + 3f\left(x + \frac{h}{3}\right) + \right.$$

$$\left. + 3f\left(x + \frac{2}{3}h\right) + f\,(x+h)\right) = -\frac{f^{(4)}\,(x)\,h^5}{6480} + O\,(h^6).$$

La comparaison des parties principales de ρ_1 et ρ_2 montre que dans notre cas, c'est-à-dire pour des équations où $f_y = 0$, il faut donner la préférence à la formule (15). Elle est donc à préférer également dans le cas d'équations avec f_y proche de zéro.

D'ordinaire, on donne la préférence à la formule (14). Les traditions en calcul numérique appellent à vrai dire une attitude double. En effet, si une méthode de la classe considérée convient à une certaine époque, les utilisateurs s'y habituent avec le temps, si bien que l'abandonner pour une autre, fût-elle plus efficace, exige une « rééducation psychologique » et un certain temps pour « se faire » à la technique nouvelle. Pour que cette technique soit acceptée par la majorité, il faut qu'elle présente des avantages sérieux pour une caractéristique quelconque. Nous en avons déjà parlé dans le chapitre VII, lorsque nous avons discuté la mise en pratique des méthodes mathématiques de résolution des problèmes économiques.

La refonte indiquée s'échelonne parfois sur une longue période. Les méthodes numériques, ainsi que les procédés de fabrication, tendent à la perfection et on peut parler de leur raffinement en moyenne et par unité de temps (par an, par exemple). Si la refonte envisagée fournit un effet mineur et qu'elle entrave la mise en service de méthodes et de procédés de fabrication nouveaux, il est préférable de s'en abstenir. Tout comme dans la planification, le passage à des techniques nouvelles peut être cause de difficultés absentes de l'analyse théorique primitive. Il se peut que la méthode combine une erreur par pas moindre avec une erreur globale plus grosse ou avec une sensibilité plus grande à l'erreur de calcul. Entre les procédés (6), (7), (12), nous aurions choisi le dernier en nous appuyant sur les considérations ci-dessus (douteuses à vrai dire) sur l'espérance mathématique du carré de l'erreur dans ces formules. Or, lorsqu'il s'agit de procédés d'intégration exacts au second ordre, l'usage accorde la préférence à la méthode (6) où l'espérance mathématique du carré de la partie principale de l'erreur est plus importante.

En comparant les deux formules exactes au quatrième ordre, on opte pour la formule (15) avec hf_y faible. D'aucuns estiment que celle-ci n'est avantageuse que pour hf_y très petits et que si l'intégration est effectuée avec une précision modérée, on choisit la formule (14). La notion de pratique du calcul numérique est en fait assez floue. Le nombre de classes d'équations différentielles réelles est sensiblement supérieur à celui des problèmes sur lesquels on compare les méthodes d'intégration numérique; aussi les jugements basés sur la « pratique » ne sont-ils pas toujours objectifs. Le critère de pratique a pourtant beau être vague, il renferme souvent une information positive impossible à formaliser et à justifier à une étape donnée du développement de la science. On a certaines raisons mathématiques d'opter pour les formules (14) avec $f_y < 0$ et des solutions régulières du moment qu'elles permettent d'effectuer une intégration avec un pas plus grand encore que la méthode reste peu sensible à l'erreur de calcul; notons néanmoins que ces raisons ont été découvertes après que la pratique de l'intégration numérique d'équations différentielles ordinaires a arrêté son choix sur les formules (14).

Dans une situation analogue à celle que nous venons d'examiner on n'arrive généralement pas à proposer une solution complètement justifiée (du point de vue mathématique) du problème. Dans notre cas s'y ajoute encore l'impossibilité de décrire de façon satisfaisante la classe d'équations différentielles à résoudre.

Il sera fondamental pour l'exposé ultérieur que l'erreur de méthode par pas $\varphi(h)$ possède une partie principale, i. e.

$$\varphi(h) = \psi(x, y)\, h^{s+1} + O(h^{s+2}). \tag{16}$$

Esquissons les grandes lignes de la démonstration. Supposons que le deuxième membre et toutes ses dérivées jusqu'à l'ordre $s+1$ compris sont uniformément bornés dans un domaine G:

$$x_0 \leqslant x \leqslant x_0 + X, \quad |y| < \infty.$$

Il en sera alors de même des dérivées jusqu'à l'ordre $s + 2$ inclusivement de toutes les solutions de l'équation. D'après la formule de Taylor, l'égalité (8) se met sous une forme plus précise

$$\varphi(h) = \frac{\varphi^{(s+1)}(0)}{(s+1)!} h^{s+1} + \frac{\varphi^{(s+2)}(\theta h)}{(s+2)!} h^{s+2}.$$

On a

$$\varphi^{(s+1)}(0) = y^{(s+1)}(0) - z^{(s+1)}(0).$$

On s'assure sans peine que $y^{(s+1)}(0)$ et $z^{(s+1)}(0)$ s'explicitent en fonction des valeurs en (x, y) de la fonction f et de ses dérivées d'ordre s au plus (nous avons déjà obtenu de telles expressions explicites).

Comme le deuxième membre est $s + 1$ fois dérivable, il en découle que la fonction $\psi(x, y)$ est dérivable dans le domaine G et que ses dérivées ψ_x et ψ_y sont uniformément bornées dans G. On établit de façon analogue que $\varphi^{(s+2)}(\theta h)$ est uniformément borné pour $x_0 \leqslant x < x + h \leqslant x_0 + X$. Certains cas particuliers de cette affirmation découlent également des relations écrites plus haut.

§ 3. Méthodes avec contrôle de l'erreur par pas

L'optimisation d'un algorithme d'intégration numérique exigeait (ch. III) que les points soient répartis de façon que l'intégration sur chaque segment élémentaire soit entachée d'une erreur sensiblement identique; à cette fin on doit avoir une idée de l'ordre de grandeur de l'erreur sur chaque segment partiel. On retrouve exactement le même problème dans l'intégration numérique d'équations différentielles ordinaires.

Nous allons raisonner de façon suivante. La partie principale de l'erreur sur un pas d'intégration est

$$\frac{\varphi^{(s+1)}(0)\, h^{s+1}}{(s+1)!}.$$

Le point $(x + h, z(h))$ est proche du point (x, y), et l'erreur sur le pas suivant renfermera la même partie principale. Au bout de deux pas on a une approximation $y^{(1)}$ de $y(x + 2h)$ telle que

$$y^{(1)} - y(x+2h) \sim 2\, \frac{\varphi^{(s+1)}(0)\, h^{s+1}}{(s+1)!}.$$

En prenant le point (x, y) pour point de départ et en appliquant la méthode de Runge-Kutta avec le pas $2h$ on obtient une valeur approchée $y^{(2)}$ telle que

$$y^{(2)} - y(x+2h) \sim \frac{\varphi^{(s+1)}(0)\, (2h)^{s+1}}{(s+1)!}.$$

On représente à l'aide de ces relations la partie principale de l'erreur par pas

$$y^{(1)} - y(x+2h) \sim \frac{y^{(2)} - y^{(1)}}{2^s - 1}.$$

On améliore $y^{(1)}$ en y ajoutant la valeur de la partie principale de l'erreur, i. e. on pose

$$y(x+2h) \approx y^{(1)} + \frac{y^{(1)} - y^{(2)}}{2^s - 1}. \tag{1}$$

Nous avons donc la possibilité d'obtenir tous les deux pas, en plus de la valeur approchée $y^{(1)}$, l'expression de la partie principale de l'erreur qu'elle détermine. Si le comportement de la solution varie brusquement, le calcul de $y^{(1)} - y^{(2)}$ met souvent en évidence une erreur très importante par pas, ce qui oblige à rejeter une partie des résultats et à changer de pas.

La recherche du pas d'intégration peut être plus souple si le pas et l'évaluation d'erreur exigent un nombre moindre de valeurs des seconds membres.

Citons un exemple de formules fournissant les mêmes caractéristiques de précision pour un recours minimal au second membre.

$$\begin{aligned}
&k_1 = hf(x, y), \quad k_2 = hf\left(x + \frac{h}{2},\ y + \frac{k_1}{2}\right),\\
&k_3 = hf\left(x + \frac{h}{2},\ y + \frac{1}{4}(k_1 + k_2)\right), \quad k_4 = hf(x+h,\ y - k_2 + 2k_3),\\
&k_5 = hf\left(x + \frac{2h}{3},\ y + \frac{1}{27}(7k_1 + 10k_3 + k_4)\right),\\
&k_6 = hf\left(x + \frac{h}{5},\ y + \frac{1}{625}(28k_1 - 125k_2 + 546k_3 + 54k_4 - 378k_5)\right),\\
&\Delta y = \frac{1}{6}(k_1 + 4k_3 + k_4);
\end{aligned} \tag{2}$$

la partie principale de l'erreur est

$$y(x+h) - z(h) = r + O(h^6),$$

$$r = \frac{1}{336}(42k_1 + 224k_3 + 21k_4 - 162k_5 - 125k_6).$$

La valeur $z(h)$ se calcule par l'une des formules de Runge-Kutta citées pp. 438-439.

En posant $y(x_0 + h) \approx z(h) + r$, on obtient une méthode du type considéré avec $s = 5$ et, partant, une erreur par pas de l'ordre de $O(h^6)$; l'utilisation directe (nous l'avons décrite plus haut) de la méthode de Runge-Kutta et l'évaluation de l'erreur par $y^{(1)} - y^{(2)}$ demandent non pas 6 mais 11 calculs du second membre.

§ 4. Evaluation d'erreur dans les méthodes à pas séparés

Considérons diverses méthodes d'intégration dans lesquelles on calcule de proche en proche les approximations y_j des valeurs $y(x_j)$, $x_0 < x_1 < \ldots < x_N = x_0 + X$. Soit k fixe et supposons

que pour $j \geqslant k$ les valeurs y_j se déterminent comme les valeurs d'une fonctionnelle

$$y_j = F(f;\ x_j,\ \ldots,\ x_{j-k};\ y_{j-1},\ \ldots,\ y_{j-k}). \tag{1}$$

On dit d'une telle méthode d'intégration numérique qu'elle est *à pas liés*. Tous les procédés d'intégration rencontrés plus haut possèdent une propriété commune : la valeur approchée de la solution en un point est définie uniquement en fonction de la valeur au point précédent ; les formules correspondantes ont par conséquent la forme (1) avec $k = 1$. Ce sont les méthodes *à pas séparés*.

On dirait de prime abord que la formule (3.1) ne se rapporte pas à la classe (1) avec $k = 1$. En effet, la valeur approchée de $y(x + 2h)$ est calculée à l'aide des approximations de $y(x)$ et $y(x + h)$. Mais la valeur approchée de $y(x + h)$ est elle-même fonction de $y(x)$, et la formule (3.1) s'écrit en fin de compte

$$y(x + 2h) \approx \Phi(f,\ x,\ 2h,\ y).$$

Si, lorsque $x_j = x$, on prend $x_{j+1} = x + 2h$, on aboutit à un cas particulier de la formule (1) avec $k = 1$.

Considérons une technique spéciale d'évaluation de l'erreur qui n'a cours que dans les méthodes à pas séparés.

Si $k = 1$, on connaît une écriture plus commode de (1) :

$$y_{j+1} = F(f,\ x_j,\ x_{j+1} - x_j,\ y_j). \tag{2}$$

Les valeurs approchées réelles de $y(x_j)$ sont reliées non pas par (2), mais par

$$y_{j+1} = F(f,\ x_j,\ x_{j+1} - x_j,\ y_j) + \delta_{j+1}. \tag{3}$$

La présence du terme δ_{j+1} s'explique par les causes suivantes :

1) arrondi lors du calcul ;

2) erreurs sur les valeurs du second membre $f(x, y)$, car, primo, $f(x, y)$ est une certaine approximation du second membre d'une équation différentielle réelle, et, secundo, en calcul automatique cette fonction est souvent approchée par des fonctions calculables moyennant les opérations élémentaires de la machine ;

3) dans certains cas la valeur y_{j+1} est déterminée par une équation équivalente à (1) mais non résolue explicitement par rapport à la variable y_{j+1} ; la quantité δ_{j+1} comprend alors une composante qui est déduite de la solution approchée de cette équation.

Très souvent on a l'estimation

$$|\delta_{j+1}| \leqslant C(|y_j| + 1)\, 2^{-t},$$

où t est la longueur du nombre machine et C une constante pas très élevée.

De même, la condition intiale y_0 diffère de la valeur de la solution cherchée $y(x_0)$ par suite d'une erreur due à la définition des données initiales et à l'arrondi. Soient $y(x)$ la solution cherchée

de l'équation différentielle et $y_j(x)$ les solutions vérifiant $y_j(x_j) = y_j$ (fig. 8.4.1). L'erreur $R_n = y_n - y(x_n)$ s'écrit

$$R_n = y_n(x_n) - y_0(x_n) + y_0(x_n) - y(x_n) = \sum_{j=1}^{n} (y_j(x_n) - y_{j-1}(x_n)) + (y_0(x_n) - y(x_n)). \quad (4)$$

La différence des solutions d'une équation différentielle en un point

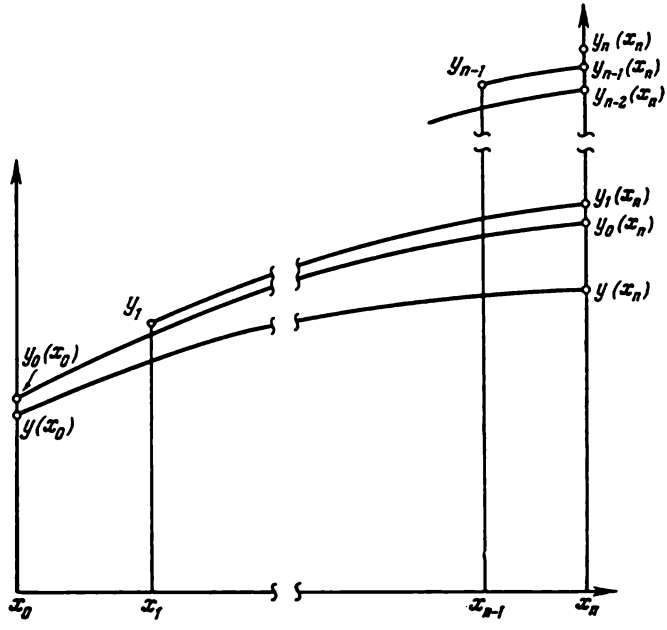


Fig. 8.4.1

peut être exprimée par leur différence en un autre. On a le

LEMME. *Soient $Y_1(x)$ et $Y_2(x)$ deux solutions de l'équation différentielle $y' = f(x, y)$. Alors*

$$Y_2(\beta) - Y_1(\beta) = (Y_2(\alpha) - Y_1(\alpha)) \exp\left(\int_{\alpha}^{\beta} f_y(x, \tilde{y}(x))\, dx\right), \quad (5)$$

avec $\tilde{y}(x)$ entre $Y_1(x)$ et $Y_2(x)$.

DÉMONSTRATION. Retranchons membre à membre les égalités

$$Y_2' = f(x, Y_2), \quad Y_1' = f(x, Y_1).$$

Selon la formule de Lagrange, la différence $f(x, Y_2) - f(x, Y_1)$ peut se mettre sous la forme $f_y(x, \tilde{y})(Y_2 - Y_1)$ avec $\tilde{y}$ compris

entre Y_1 et Y_2. On obtient une équation différentielle linéaire en $Y_2 - Y_1$:

$$(Y_2 - Y_1)' = f_y(x, \tilde{y})(Y_2 - Y_1),$$

d'où

$$(\ln(Y_2 - Y_1))' = f_y(x, \tilde{y}).$$

En substituant les limites α et β et en se débarrassant des logarithmes, on trouve la relation (5) cherchée. La fonction

$$f_y(x, \tilde{y}(x)) = \frac{f(x, Y_2(x)) - f(x, Y_1(x))}{Y_2(x) - Y_1(x)}$$

est continue si $f(x, y)$ l'est, ce qui légitime toutes les transformations ci-dessus.

Soit $\alpha = x_j$, $\beta = x_n$,

$$Y_1(x) = y_{j-1}(x), \quad Y_2(x) = y_j(x),$$

alors

$$y_j(x_n) - y_{j-1}(x_n) = (y_j(x_j) - y_{j-1}(x_j)) \exp\left(\int_{x_j}^{x_n} f_y(x, \tilde{y}_j(x))\, dx\right),$$

avec $\tilde{y}_j(x)$ entre $y_{j-1}(x)$ et $y_j(x)$.
De même

$$y_0(x_n) - y(x_n) = (y_0(x_0) - y(x_0)) \exp\left(\int_{x_0}^{x_n} f_y(x, \tilde{y}_0(x))\, dx\right).$$

L'égalité (4) s'écrit alors

$$R_n = \sum_{j=1}^{n} \omega_j \exp\left(\int_{x_j}^{x_n} f_y(x, \tilde{y}_j(x))\, dx\right) + R_0 \exp\left(\int_{x_0}^{x_n} f_y(x, \tilde{y}_0(x))\, dx\right), \quad (6)$$

où $\omega_j = y_j(x_j) - y_{j-1}(x_j)$, $j = 1, \ldots$

L'égalité (3) entraîne

$$\omega_j = y_j - y_{j-1}(x_j) = \rho_j + \delta_j,$$

avec

$$\rho_j = F(f, x_{j-1}, x_j - x_{j-1}, y_{j-1}) - y_{j-1}(x_j).$$

Voyons ce que représente la quantité ρ_j. $F(f, x_{j-1}, x_j - x_{j-1}, y_{j-1})$ est un nombre obtenu à l'aide de la formule (2) et $y_{j-1}(x_j)$ la valeur en x_j de la solution exacte de l'équation différentielle dans la condition $y_{j-1}(x_{j-1}) = y_{j-1}$. Donc, ρ_j est l'erreur par pas de la méthode considérée si on prend pour point de départ le point (x_{j-1}, y_{j-1}), s'il n'y a pas d'arrondi et si le pas est $x_j - x_{j-1}$.

Supposons que tous les j correspondant au segment d'intégration considéré $x_0 < x_j \leqslant x_0 + X$ vérifient

$$|\rho_j| \leqslant C_1 (x_j - x_{j-1})^{s+1}. \quad (7)$$

Soit

$$L = \sup_{x_0 \leqslant x \leqslant x_0 + X} |f_y| < \infty$$

et

$$H = \max_{0 < j \leqslant N} (x_j - x_{j-1}).$$

Rendant (7) moins précise, il vient

$$|\rho_j| \leqslant C_1 H^s (x_j - x_{j-1}). \tag{8}$$

Lorsque $x_0 \leqslant x_j \leqslant x_n \leqslant x_0 + X$, on a

$$\exp\Big(\int_{x_j}^{x_n} f_y(x, \tilde{y}_j(x))\,dx\Big) \leqslant \exp(L(x_n - x_j)) \leqslant \exp(LX).$$

Evaluons par cette inégalité le second membre de (6) ; il vient

$$|R_n| \leqslant \exp(LX)\Big(\sum_{j=1}^{n} (|\rho_j| + |\delta_j|) + |R_0|\Big).$$

Servons-nous maintenant de l'estimation (8) :

$$|R_n| \leqslant \exp(LX)\Big(\sum_{j=1}^{n} (C_1 H^s (x_j - x_{j-1}) + |\delta_j|) + |R_0|\Big) \leqslant$$
$$\leqslant \exp(LX)\,(C_1 H^s (x_n - x_0) + n\delta + |R_0|) \leqslant$$
$$\leqslant \exp(LX)\,(C_1 X H^s + N\delta + |R_0|) ; \tag{9}$$

$\delta = \max_j |\delta_j|$. Il en résulte

$$\max_{x_0 < x_n \leqslant x_0 + X} |R_n| \to 0 \text{ lorsque } H \to 0,$$

si $N\delta$ tend vers zéro avec $|R_0|$. Nous avons montré qu'avec un pas d'intégration suffisamment petit et une erreur de calcul faible, la solution approchée obtenue par la méthode de Runge-Kutta avoisine la solution exacte.

Dans certains cas on recherche la solution de l'équation dans un intervalle important. Les estimations d'erreur contiennent alors un très grand facteur : $\exp(LX)$. Aussi, si LX est élevé, il se peut que la précision désirée soit atteinte avec des pas si faibles et une erreur de calcul par pas tellement insignifiante qu'on est fort tenté de refuser la méthode en question. Il ne suffit donc pas de caractériser les méthodes d'après le critère de convergence de la solution approchée vers la solution exacte avec fractionnement du pas et une diminution suffisamment rapide de l'erreur.

Si $f_y(x, y) \leqslant -b < 0$, on peut éliminer dans (9) le facteur croissant brusquement avec X. Examinons le cas d'un

pas constant $x_j - x_{j-1} \equiv h$. Alors

$$\exp\Big(\int_{x_j}^{x_n} f_y(x, \tilde{y}(x))\,dx\Big) \leqslant \exp(-b(x_n - x_j)) \leqslant \exp(-b(n-j)h);$$

soit

$$|\omega_j| \leqslant C_1 h^{s+1} + \delta.$$

Evaluons le second membre de (6), il vient

$$|R_n| \leqslant \sum_{j=1}^{n} (C_1 h^{s+1} + \delta) \exp(-b(n-j)h) + |R_0| \exp(-bnh). \qquad (10)$$

On a

$$\sum_{j=1}^{n} \exp(-b(n-j)h) \leqslant \sum_{k=0}^{\infty} \exp(-bkh) = \frac{1}{1-\exp(-bh)}.$$

L'estimation d'erreur s'écrit finalement

$$|R_n| \leqslant \frac{C_1 h^{s+1} + \delta}{1-\exp(-bh)} + |R_0| \exp(-bnh). \qquad (11)$$

Puisque $1 - \exp(-bh) \sim |b|h$,

$$|R_n| \leqslant C_2 \left(h^s + \frac{\delta}{h}\right) + |R_0| \exp(-bnh).$$

Formellement cette estimation ne dépend pas de la longueur X de l'intervalle d'intégration, mais celle-ci peut influer de facon indirecte sur la valeur de C_2 par intermédiaire des estimations des dérivées (nous en reparlerons au § 12).

PROBLÈME. Supposons qu'une fonction f et toutes ses dérivées jusqu'à l'ordre $s+1$ inclusivement sont uniformément bornées dans un domaine G, et soient

$$|R_0| = o(h^s) \quad \text{et} \quad \delta = \max_{x_0 \leqslant x_j \leqslant x_0 + X} |\delta_j| = o(h^s).$$

Démontrer qu'en intégrant par la méthode de Runge-Kutta exacte à l'ordre s on a

$$\max_{x_0 \leqslant x_n \leqslant x_0 + X} \left| R_n - h^s \int_{x_0}^{x_n} \psi(x, y(x)) \exp\Big(\int_{x}^{x_n} f_y(t, y(t))\,dt\Big)\,dx \right| = o(h^s); \qquad (12)$$

pour la définition de la fonction $\psi(x, y)$ voir (2.16). La relation (12) donne une idée de la partie principale de l'erreur dans la méthode de Runge-Kutta.

§ 5. Méthodes des différences finies

Parmi les méthodes à pas liés, les plus répandues sont les méthodes d'intégration sur un réseau de pas constant $x_j - x_{j-1} \equiv h = \text{const}$ moyennant des relations de la forme

$$\sum_{i=0}^{k} a_{-i} y_{j-i} - h \sum_{i=0}^{k} b_{-i}(x_{j-i}, y_{j-i}) = 0; \tag{1}$$

a_i est constant et

$$\left|\frac{\partial b_{-i}}{\partial y_{j-i}}\right| \leqslant C \quad \text{pour } h \leqslant h_0.$$

De tous ces procédés on préfère généralement

$$\sum_{i=0}^{k} a_{-i} y_{j-i} - h \sum_{i=0}^{k} b_{-i} f(x_{j-i}, y_{j-i}) = 0, \tag{2}$$

qu'on appelle *méthodes des différences finies* ou *schémas aux différences finies.*

Les méthodes (2) qui sont à la fois les plus simples et les plus utilisées se déduisent de formules de quadrature; toute formule

$$\int_{-ph}^{0} f(x)\,dx \approx h \sum_{i=0}^{m} b_{-i} f(-ih), \quad m \geqslant 0, \tag{3}$$

engendre une formule correspondante d'intégration numérique d'équations différentielles ordinaires. En effet, le report de $f(x) = y'(x_j + x)$ dans (3) donne

$$\int_{-ph}^{0} f(x)\,dx = y(x_j) - y(x_{j-p}) \approx h \sum_{i=0}^{m} b_{-i} y'(x_{j-i}),$$

et la substitution de $f(x, y(x))$ à $y'(x)$

$$y(x_j) - y(x_{j-p}) \approx h \sum_{i=0}^{m} b_{-i} f(x_{j-i}, y(x_{j-i})). \tag{4}$$

Le schéma aux différences finies correspondant s'écrit

$$y_j - y_{j-p} - h \sum_{i=0}^{m} b_{-i} f(x_{j-i}, y_{j-i}) = 0. \tag{5}$$

Dans la pratique on se sert des formules (1), (2) avec $a_0 \neq 0$, $b_0 = 0$, appelées *formules d'extrapolation* ou *explicites*, et des mêmes formules avec $a_0 \neq 0$, $b_0 \neq 0$, dites *d'interpolation* ou *implicites.*

Dans la suite nous supposerons toujours $a_0 \neq 0$.

Citons plusieurs formules (5) et pour abréger désignons

$$f_n = f(x_n, y_n).$$

La formule implicite

$$y_j - y_{j-1} = \frac{h}{2}(f_j + f_{j-i}) \tag{6}$$

correspond à la formule des trapèzes

$$\int_{-h}^{0} f(x)\,dx \approx \frac{h}{2}(f(0) + f(-h));$$

de même on a

$$y_j - y_{j-2} = \frac{h}{3}(f_j + 4f_{j-1} + f_{j-2}) \tag{7}$$

pour la formule de Simpson

$$\int_{-2h}^{0} f(x)\,dx \approx \frac{h}{3}(f(0) + 4f(-h) + f(-2h))$$

et

$$y_j - y_{j-3} = \frac{h}{8}(3f_j + 9f_{j-1} + 9f_{j-2} + 3f_{j-3}) \tag{8}$$

pour

$$\int_{-3h}^{0} f(x)\,dx \approx \frac{h}{8}(3f(0) + 9f(-h) + 9f(-2h) + 3f(-3h)).$$

La formule des rectangles

$$\int_{-h}^{0} f(x)\,dx \approx hf\left(-\frac{h}{2}\right)$$

diffère de (3). Cependant, en remplaçant h par $2h$, on obtient la formule de quadrature

$$\int_{-2h}^{0} f(x)\,dx \approx 2hf(-h)$$

et la formule explicite correspondante

$$y_j - y_{j-2} = 2hf_{j-1}. \tag{9}$$

Si le reste de (3) est évalué par $D(q) \max |f^{(q)}(x)| h^{q+1}$, l'erreur résultant de l'égalité (4) l'est par $D(q) \max |y^{(q+1)}(x)| h^{q+1}$.

Parmi les formules (5) les plus usitées sont celles qui correspondent à $p = 1$. On les appelle *formules d'Adams*. Construisons-les

et évaluons leur reste. Selon la formule de Newton,

$$f(x) = Q_m(x) + r_m(x),$$

où $Q_m(x)$ est le polynôme d'interpolation basé sur les points $(0, \ldots, -mh)$

$$Q_m(x) = \sum_{i=0}^{m} \frac{\nabla^i f(0)}{h^i}\, \omega_{hi}(x), \quad \omega_{h0} = 1, \tag{10}$$

$$\omega_{hi} = \frac{x(x+h)\ldots(x+(i-1)h)}{i!} \quad \text{pour} \quad i > 0,$$

$$r_m(x) = f(0), -h, \ldots, -mh, x)(m+1)!\,\omega_{h,\,m+1}.$$

Corrélativement, on a

$$\int_0^h f(x)\,dx = \int_0^h Q_m(x)\,dx + \int_0^h r_m(x)\,dx. \tag{11}$$

La fonction $\omega_{h,\,m+1}(x)$ garde son signe sur le segment $[0, h]$. Le théorème de la moyenne nous autorise donc à écrire

$$\int_0^h r_m(x)\,dx = f(0;\ -h;\ \ldots;\ -mh;\ \zeta)(m+1)! \int_0^h \omega_{h,\,m+1}(x)\,dx$$

avec $0 \leqslant \zeta \leqslant h$. Le changement de variable $x = ht$ dans $\int_0^h \omega_{hi}(x)\,dx$ fournit

$$\int_0^h \omega_{hi}(x)\,dx = \gamma_i h^{i+1}, \tag{12}$$

où

$$\gamma_i = \int_0^1 \frac{t(t+1)\ldots(t+i-1)}{i!}\,dt.$$

Vu ces relations l'égalité (11) se récrit:

$$\int_0^h f(x)\,dx = h\sum_{i=0}^{m} \gamma_i \nabla^i f(0) + \\ + \gamma_{m+1} h^{m+2} (m+1)!\, f(0;\ -h;\ \ldots;\ -mh;\ \zeta).$$

En écrivant au lieu de $(m+1)!\,f(0;\ -h;\ \ldots;\ -mh;\ \zeta)$ la valeur de la dérivée au point médian, on trouve, conformément à (2.5.4),

$$\int_0^h f(x)\,dx = h\sum_{i=0}^{m} \gamma_i \nabla^i f(0) + \gamma_{m+1} h^{m+2} f^{(m+1)}(\eta) \tag{13}$$

avec $-mh \leqslant \eta \leqslant h$. Il en résulte, après la substitution $f(x) = y'(x_j + x - h)$ et l'élimination du reste, la formule de calcul correspondante

$$y_j - y_{j-1} = h \sum_{i=0}^{m} \gamma_i \nabla^i f_{j-1}. \tag{14}$$

Pour obtenir les formules d'Adams implicites prenons un support d'interpolation formé des points h; 0; ...; $-(m-1)h$; il vient

$$f(x) = \overline{Q}_m(x) + \overline{r}_m(x),$$

où

$$\overline{Q}_m(x) = \sum_{i=0}^{m} \frac{\nabla^i f(h)}{h^i} \overline{\omega}_{h^i}(x),$$

$$\overline{\omega}_{h^i}(x) = \frac{(x-h)\,x\,(x+h)\ldots(x+(i-2)\,h)}{i!},$$

$$\overline{r}_m(x) = f(h;\ 0;\ \ldots;\ -(m-1)\,h;\ x)\,(m+1)!\,\overline{\omega}_{h,\,m+1}(x).$$

Moyennant des transformations analogues, on a

$$\int_0^h f(x)\,dx = h \sum_{i=0}^{m} \overline{\gamma}_i \nabla^i f(h) + \overline{\gamma}_{m+1} h^{m+2} f^{(m+1)}(\eta), \tag{15}$$

avec

$$-(m-1)\,h \leqslant \eta \leqslant h, \quad \overline{\gamma}_i = \int_0^1 \frac{(t-1)\,t\,(t+1)\ldots(t+i-2)}{i!}\,dt.$$

La formule de calcul correspondante a la forme

$$y_j - y_{j-1} = h \sum_{i=0}^{m} \overline{\gamma}_i \nabla^i f_j. \tag{16}$$

Voyons le comportement des coefficients γ_i et $\overline{\gamma}_i$.

Evidemment, γ_0, $\overline{\gamma}_0 > 0$ et $\gamma_i > 0$, $\overline{\gamma}_i < 0$, lorsque $i > 0$. Avec le changement de variable $t = 1 - u$ les coefficients γ_i s'écrivent

$$\gamma_i = \int_0^1 \frac{(1-u)\ldots(i-u)}{i!}\,du = \int_0^1 \prod_{k=1}^{i} \left(1 - \frac{u}{k}\right) du.$$

On a

$$1 - \frac{u}{k} = \exp\left(\ln\left(1 - \frac{u}{k}\right)\right) = \exp\left(-\frac{u}{k} + \varepsilon_k\right),$$

où

$$0 \leqslant \varepsilon_k < \frac{u^2}{2k^2}.$$

Donc,

$$\prod_{k=1}^{i}\left(1-\frac{u}{k}\right)=\exp\left(-s(i)\,u\right)\exp\left(\sum_{k=1}^{i}\varepsilon_k\right), \qquad s(i)=\sum_{k=1}^{i}\frac{1}{k}.$$

On a évidemment

$$0 \leqslant \sum_{k=1}^{i}\varepsilon_k < \sum_{k=1}^{\infty}\frac{1}{2k^2}=\frac{\pi^2}{12}.$$

On a finalement l'estimation

$$v_i(u) \leqslant \prod_{k=1}^{i}\left(1-\frac{u}{k}\right) \leqslant \exp\left(\frac{\pi^2}{12}\right) v_i(u),$$

$v_i(u)=\exp\left(-s(i)\,u\right)$ et, par conséquent,

$$\int_0^1 v_i(u)\,du \leqslant \gamma_i \leqslant \exp\left(\frac{\pi^2}{12}\right)\int_0^1 v_i(u)\,du. \tag{17}$$

Du fait que $s(i) \sim \ln i$,

$$\int_0^1 v_i(u)\,du=\frac{1-\exp\left(-s(i)\right)}{s(i)} \sim \frac{1}{\ln i}.$$

Cette dernière relation et (17) entraînent que γ_i est de l'ordre de $\frac{1}{\ln i}$.

On a l'identité évidente

$$(t-1)\,t\,\ldots\,(t+i-2)+it\,(t+1)\,\ldots\,(t+i-2)= \\ =t\,\ldots\,(t+i-1).$$

Divisons-la par $i!$ et intégrons par rapport à t entre 0 et 1 ; il vient

$$\bar{\gamma}_i+\gamma_{i-1}=\gamma_i;$$

comme tous les $\gamma_i > 0$ et $\bar{\gamma}_i < 0$ lorsque $i > 0$, on en conclut que les γ_i décroissent de façon monotone et tous les $|\bar{\gamma}_i| < \gamma_i$.

Ainsi, le facteur constant dans l'estimation de l'erreur par pas est toujours plus petit dans les méthodes d'Adams implicites que dans les méthodes explicites exactes au même ordre.

Citons quelques valeurs à titre d'exemple

$$\{\gamma_p\}=1,\ \frac{1}{2},\ \frac{5}{12},\ \frac{3}{8},\ \frac{251}{720},\ \ldots,\quad \{\bar{\gamma}_p\}=1,\ -\frac{1}{2},\ -\frac{1}{12},\ -\frac{1}{24},\ \ldots$$

PROBLÈME. Montrer que $\gamma_i \sim \frac{\text{const}}{\ln i}$, $\bar{\gamma}_i \sim \frac{\text{const}}{i \ln i}$, $i \to \infty$.

§ 6. Méthode des coefficients indéterminés

On construit également les formules d'intégration numérique par la *méthode des coefficients indéterminés*. Approchons la dérivée $y'(x_j)$ par

$$\sum_{i=0}^{k} \frac{a_{-i} y(x_{j-i})}{h}$$

et $f(x_j, y(x_j))$ par

$$\sum_{i=0}^{k} b_{-i} f(x_{j-i}, y(x_{j-i}))$$

(a_{-i} et b_{-i} sont supposés indépendants de h). D'où l'égalité approchée

$$\sum_{i=0}^{k} \frac{a_{-i} y(x_{j-i})}{h} \approx \sum_{i=0}^{k} b_{-i} f(x_{j-i}, y(x_{j-i})) \tag{1}$$

à laquelle correspond le schéma aux différences finies

$$\sum_{i=0}^{k} \frac{a_{-i} y_{j-i}}{h} - \sum_{i=0}^{k} b_{-i} f_{j-i} = 0. \tag{2}$$

La quantité

$$r_j = \sum_{i=0}^{k} \frac{a_{-i} y(x_{j-i})}{h} - \sum_{i=0}^{k} b_{-i} f(x_{j-i}, y(x_{j-i}))$$

s'appelle *l'erreur d'approximation de l'équation différentielle initiale par le schéma* (2).

DÉFINITION. On dit qu'*un problème aux différences approxime un problème différentiel* sur le segment $[x_0, x_0+X]$ si

$$\max_{x_0 \leqslant x_j \leqslant x_0+X} |r_j| \to 0 \quad \text{pour} \quad h \to 0.$$

En se rappelant que $f(x_{j-i}, y(x_{j-i})) = y'(x_{j-i})$ et $x_{j-i} = x_j - ih$, on obtient

$$r_j = \sum_{i=0}^{k} \frac{a_{-i} y(x_j - ih)}{h} - \sum_{i=0}^{k} b_{-i} y'(x_j - ih).$$

Supposons que toutes les dérivées de la solution jusqu'à l'ordre q inclus sont bornées :

$$|y^{(p)}(x)| \leqslant M_p < \infty \quad \text{pour} \quad x_0 \leqslant x \leqslant x_0+X, \quad p = 0, \ldots, q.$$

Représentons toutes les quantités

$$y(x_j - ih) \quad \text{et} \quad y'(x_j - ih)$$

à l'aide de la formule de Taylor:

$$y(x_j - ih) = \sum_{p=0}^{q-1} y^{(p)}(x_j) \frac{(-ih)^p}{p!} + \beta_j^i,$$

$$y'(x_j - ih) = \sum_{p=1}^{q-1} y^{(p)}(x_j) \frac{(-ih)^{p-1}}{(p-1)!} + \gamma_j^i;$$

en conformité avec l'évaluation du reste de la série taylorienne, on a

$$|\beta_j^i| \leqslant M_q (ih)^q / q!,$$
$$|\gamma_j^i| \leqslant M_q (ih)^{q-1} / (q-1)!$$

Portons les expressions de $y(x_j - ih)$ et $y'(x_j - ih)$ dans le second membre de la représentation de r_j et regroupons les coefficients de $y^{(p)}(x_j)$; nous trouvons

$$r_j = E_0 h^{-1} y(x_j) + E_1 y'(x_j) + \ldots + E_{q-1} h^{q-2} y^{(q-1)}(x_j) + \varepsilon_j, \qquad (3)$$

avec

$$E_0 = \sum_{i=0}^{k} a_{-i},$$

$$E_p = \sum_{i=0}^{k} \frac{a_{-i}(-i)^p}{p!} - \sum_{i=0}^{k} \frac{b_{-i}(-i)^{p-1}}{(p-1)!} \qquad (4)$$

lorsque $p > 0$, et

$$\varepsilon_j = \sum_{i=0}^{k} \frac{a_{-i}\beta_j^i}{h} - \sum_{i=0}^{k} b_{-i}\gamma_j^i.$$

Evidemment,

$$|\varepsilon_j| \leqslant D_q M_q h^{q-1},$$

où

$$D_q = \sum_{i=0}^{k} |a_{-i}| \frac{i^q}{q!} + \sum_{i=0}^{k} |b_{-i}| \frac{i^{q-1}}{(q-1)!}.$$

Dans certains cas (les méthodes d'Adams du § 5 par exemple), on réussit à réduire la valeur de la constante D_q dans l'estimation $|\varepsilon_j|$. Si $E_0 = \ldots = E_m = 0$, alors $r_j = O(h^m)$ et on dit que (2) est un schéma *approché à l'ordre m*. On conçoit que tout schéma approché à l'ordre m est approché à l'ordre q si $q < m$. Si $E_0 = \ldots = E_m = 0$ et $E_{m+1} \neq 0$, alors *l'ordre d'approximation est strictement égal à m*. Lorsque (2) est multiplié par une constante arbitraire, r_j se multiplie par cette même constante. Aussi la définition

de l'erreur d'approximation que nous avons adoptée n'est pas univoque. Elle le deviendra sous les conditions supplémentaires suivantes : quelle que soit $y(x)$ régulière,

$$\lim_{h\to 0} \sum_{i=0}^{k} \frac{a_{-i} y(x-ih)}{h} = y'(x),$$
$$\lim_{h\to 0} \sum_{i=0}^{k} b_{-i} f(x-ih,\ y(x-ih)) = f(x,\ y(x)). \tag{5}$$

LEMME. *Les conditions* (5) *sont vérifiées si et seulement si*

$$E_0 = E_1 = 0, \quad b_0 + \ldots + b_{-k} = 1. \tag{6}$$

Développons les expressions sous le signe limite en série de Taylor au point x en retenant dans le premier cas les termes de l'ordre de 1 et ceux en h dans le deuxième. On a

$$\lim_{h\to 0} \Big(\Big(\sum_{i=0}^{k} \frac{a_{-i}}{h} y(x) \Big) + \sum_{i=0}^{k} a_{-i} (-i) y'(x) + O(h) \Big) = y'(x),$$
$$\lim_{h\to\infty} \Big(\Big(\sum_{i=0}^{k} b_{-i} \Big) f(x, y(x)) + O(h) \Big) = f(x,\ y(x)).$$

Pour que ces relations soient satisfaites, il faut et il suffit que

$$\sum_{i=0}^{k} a_{-i} = 0, \qquad -\sum_{i=0}^{k} i a_{-i} = 1, \qquad \sum_{i=0}^{k} b_{-i} = 1.$$

Le premier membre de la première égalité vaut E_0, la différence des membres homologues de la deuxième et de la troisième est E_1. D'où le lemme.

Si $E_0 = E_1 = 0$, $\omega = b_0 + \ldots + b_{-k} \neq 0, 1$, une multiplication de (2) par ω^{-1} fournit un schéma aux différences qui vérifie les conditions (6) et donc les conditions de normalisation (5). Dans la suite on supposera que les conditions (5), (6) sont réalisées. L'étude de schémas avec $\omega = 0$ ne présente aucun intérêt pratique.

PROBLÈME. Démontrer que la condition $E_0 = E_1 = 0$ est nécessaire pour que la solution du problème discret converge vers celle du problème différentiel.

Les équations $E_0 = \ldots = E_m = 0$ forment un système linéaire homogène d'équations algébriques par rapport à $2k + 2$ inconnues. Si l'on a plus d'inconnues que d'équations : $2k + 2 > m + 1$ (ou, ce qui revient au même, $2k \geqslant m$), le système admet une solution non nulle. En résolvant le système d'équations $E_0 = \ldots = E_{2k} = 0$, on aboutit donc à un schéma approché à l'ordre $2k$. On a parfois

besoin d'approximations avec $b_0 = 0$. Le système d'équations $b_0 = \ldots = 0$, $E_0 = \ldots = E_{2k-1} = 0$ renferme plus d'inconnues que d'équations; il a par conséquent une solution non triviale à laquelle correspond un schéma approché à l'ordre $2k - 1$.

Exemples. Lorsque $k = 2$, la solution générale du système d'équations $E_0 = \ldots = E_4 = 0$ s'écrit $a_0 = c$, $a_{-1} = 0$, $a_{-2} = -c$, $b_0 = b_{-2} = \frac{c}{3}$, $b_{-1} = \frac{4}{3}c$. La condition de normalisation (6) donne $c = \frac{1}{2}$; le schéma de calcul revêt la forme

$$\frac{y_j - y_{j-2}}{2h} - \left(\frac{1}{6} f_j + \frac{2}{3} f_{j-1} + \frac{1}{6} f_{j-2}\right) = 0.$$

Notons que nous l'avons déjà obtenu à partir de la formule de Simpson.

Sous les conditions $b_0 = E_0 = E_1 = E_2 = E_3 = 0$, on a le schéma

$$\frac{y_j + 4y_{j-1} - 5y_{j-2}}{6h} - \left(\frac{2}{3} f_{j-1} + \frac{1}{3} f_{j-2}\right) = 0. \tag{7}$$

Si les égalités

$$\begin{gathered} y'(x_j) \approx \sum_{i=0}^{k} \frac{a_{-i} y(x_{j-i})}{h}, \\ f(x_j, y(x_j)) \approx \sum_{i=0}^{k} b_{-i} f(x_{j-i}, y(x_{j-i})) \end{gathered} \tag{8}$$

sont exactes aux termes en $O(h^m)$ près, la quantité r_j, différence des erreurs sur ces relations, est de l'ordre de $O(h^m)$ et donc, selon (3), $E_0 = \ldots = E_m = 0$. Cependant, pour que la condition $r_j = O(h^m)$ soit remplie il n'est pas nécessaire d'exiger que l'erreur sur ces relations approchées soit de l'ordre de $O(h^m)$. Ainsi, dans notre dernier schéma $r_j = O(h^3)$ et l'erreur sur (8) est en $O(h)$.

A part les méthodes de Runge-Kutta et les schémas aux différences finies, il convient de mentionner les *méthodes de Butcher* dans lesquelles chaque y_n est recherché à l'aide de quelques valeurs précédentes y_{n-i} (comme dans les méthodes des différences finies) et où on exécute à chaque pas plusieurs calculs du second membre (comme dans les méthodes de Runge-Kutta).

Voici plusieurs formules de calcul des méthodes de Butcher:

1.

$$y_{n-\frac{1}{2}} = y_{n-2} + \frac{h}{8}(9f_{n-1} + 3f_{n-2}),$$

$$y_n^1 = \frac{1}{5}(28y_{n-1} - 23y_{n-2}) + \frac{h}{15}(32f_{n-\frac{1}{2}} - 60f_{n-1} - 26f_{n-2}),$$

$$y_n = \frac{1}{31}(32y_{n-1} - y_{n-2}) + \frac{h}{93}(64f_{n-\frac{1}{2}} + 15f_n^1 + 12f_{n-1} - f_{n-2}),$$

avec une erreur par pas de l'ordre de $O(h^6)$; ici $f_m^k = f(x_m, y_m^k)$.

2.

$$y_{n-\frac{1}{2}} = \frac{1}{128}(-225y_{n-1}+200y_{n-2}+153y_{n-3}) + \frac{h}{128}(225f_{n-1}+300f_{n-2}+45f_{n-3}),$$

$$y_n^1 = \frac{1}{31}(540y_{n-1}-297y_{n-2}-212y_{n-3}) + \\ + \frac{h}{155}(384f_{n-\frac{1}{2}}-1395f_{n-1}^1-2130f_{n-2}-309f_{n-3}),$$

$$y_n = \frac{1}{617}(783y_{n-1}-135y_{n-2}-31y_{n-3}) + \\ + \frac{h}{3085}(2304f_{n-\frac{1}{2}}+465f_n^1-135f_{n-1}-495f_{n-2}-39f_{n-3}),$$

avec une erreur par pas de l'ordre de $O(h^8)$.

§ 7. Etude des propriétés des méthodes des différences finies sur des problèmes modèles

Pour comprendre quelles doivent être les conditions pour que les approximations obtenues par une méthode des différences finies convergent vers la solution du problème, étudions quelques modèles parmi les plus simples.

Considérons le comportement de la solution du schéma aux différences finies

$$\sum_{i=0}^{k} a_{-i}y_{j-i} - h\sum_{i=0}^{k} b_{-i}f_{j-i} = 0 \qquad (1)$$

sur l'exemple de l'équation $y' = My$, $M = \text{const}$. La relation (1) prend alors la forme

$$\sum_{i=0}^{k} a_{-i}y_{j-i} - h\sum_{i=0}^{k} b_{-i}My_{j-i} = 0,$$

ou encore,

$$\sum_{i=0}^{k} (a_{-i} - Mhb_{-i})\, y_{j-i} = 0. \qquad (2)$$

Supposons ensuite que $a_0 \neq 0$ et Mh tellement petit que le coefficient $a_0 - Mhb_0$ est différent de zéro.

On sait (§ 7, chap. II) que la solution de l'équation aux différences (2) à coefficients constants s'exprime à l'aide des racines de l'équation caractéristique

$$\sum_{i=0}^{k} (a_{-i} - Mhb_{-i})\, \mu^{k-i} = 0.$$

L'écriture plus commode est

$$F(Q, \mu) = \sum_{i=0}^{k} (a_{-i} - Qb_{-i})\, \mu^{k-i} = 0 \tag{3}$$

avec $Q = Mh$. Les racines du polynôme sont des fonctions continues de ses coefficients dans le domaine où son coefficient dominant est non nul. Les coefficients $a_{-i} - Qb_{-i}$ sont également des fonctions continues de Q. Par conséquent, les racines de (3) dépendent de façon continue de Q dans le domaine où $a_0 - Qb_0 \neq 0$, en particulier pour Q petit. Nous en concluons que dans cette dernière condition les racines de (3) sont proches de celles de

$$F(0, \mu) = F(\mu) = \sum_{i=0}^{k} a_{-i}\mu^{k-i} = 0, \tag{4}$$

appelée *équation caractéristique du schéma aux différences* (1). Désignons par $\mu_i(Q)$, $i = 1, \ldots, k$, les racines de (3) et par μ_i les valeurs $\mu_i(0)$, i.e. les racines de (4).

Dans la condition $\max_i |\mu_i| > 1$ le schéma aux différences est à rejeter. En effet, la solution de l'équation aux différences s'écrit

$$\sum_{i=0}^{k} c_i (\mu_i(Q))^n$$

(on prend le cas où les racines de (3) sont simples). En arrondissant on perturbe les coefficients c_i. Si le pas d'intégration est petit, $\mu_i(Q)$ est voisin de μ_i, et n peut être important ; il y a alors, parmi les $(\mu_i(Q))^n$, des quantités très grandes. Ce qui fait que de faibles perturbations de c_i peuvent induire une perturbation catastrophique de la solution. Ci-dessous le raisonnement rigoureux correspondant.

Les quantités réelles y_j obtenues par le calcul ne vérifient pas exactement l'équation (2) à cause d'arrondis. Considérons un cas « partiellement idéalisé » où on n'arrondit pas dans l'intégration numérique selon la formule (2) mais où y_{k-1} est entachée d'une erreur.

Soient y_j^q, $q = 1, 2$ des solutions des (2) qui se confondent pour

$$j = 0, \ldots, k-2, \quad \text{et} \quad y_{k-1}^2 - y_{k-1}^1 = \delta.$$

Leur différence $z_j = y_j^2 - y_j^1$ est solution de (2) sous les conditions initiales $z_0 = \ldots = z_{k-2} = 0$, $z_{k-1} = \delta$. Notons E l'opérateur de translation suivant les nœuds du réseau défini par $Ey_j = y_{j+1}$. Evidemment, $E^p y_j = y_{j+p}$ pour p entiers positifs ou négatifs, et l'équation (2) se récrit

$$\sum_{i=0}^{k} (a_{-i} - Qb_{-i})\, E^{-i} y_j = 0.$$

Soit $\mu_l(Q)$ une racine de (3) ; alors

$$\sum_{i=0}^{k} (a_{-i} - Qb_{-i})\, \mu^{k-i} = (\mu - \mu_l(Q)) \Big(\sum_{i=0}^{k-1} c_{-i}(Q)\, \mu^{k-1-i} \Big)$$

avec $c_0 = a_0 - Qb_0$. Multiplions cette égalité par μ^{l-k} et substituons à μ l'expression de E ; on a

$$E \sum_{i=0}^{k} (a_{-i} - Qb_{-i}) E^{-i} = (E - \mu_l(Q)) \Big(\sum_{i=0}^{k-1} c_{-i}(Q) E^{-i} \Big).$$

Faisant agir les opérateurs des deux membres sur la fonction z_j, on obtient

$$0 = (E - \mu_l(Q)) \Big(\sum_{i=0}^{k-1} c_{-i}(Q) E^{-i} \Big) z_j.$$

Désignons $\Big(\sum_{i=0}^{k-1} c_{-i}(Q) E^{-i} \Big) z_j$ par v_j ; puisque $E^{-i} z_j = z_{j-i}$, il vient

$$v_{k-1} = \sum_{i=0}^{k-1} c_{-i}(Q) z_{k-1-i} = c_0(Q) z_{k-1} = c_0(Q)\, \delta.$$

La fonction v_j vérifie

$$(E - \mu_l(Q)) v_j = 0, \quad \text{ou encore} \quad v_{j+1} = \mu_l(Q) v_j.$$

Il s'ensuit

$$v_j = (\mu_l(Q))^{j-(k-1)} v_{k-1} = (\mu_l(Q))^{j-(k-1)} c_0(Q)\, \delta,$$

i.e.

$$\sum_{i=0}^{k-1} c_{-i}(Q) z_{j-i} = (\mu_l(Q))^{j-(k-1)} c_0(Q)\, \delta.$$

On a l'inégalité évidente

$$\Big| \sum_{i=0}^{k-1} c_{-i}(Q) z_{j-i} \Big| \leqslant \Big(\sum_{i=0}^{k-1} |c_{-i}(Q)| \Big) \max_{j-k<p\leqslant j} |z_p|.$$

C'est pourquoi la dernière égalité donne l'estimation

$$\max_{j-k<p\leqslant j} |z_p| \geqslant |\mu_l(Q)|^{j+1-k} \frac{|c_0(Q)|}{\sum_{i=0}^{k-1} |c_{-i}(Q)|} |\delta|.$$

Une perturbation δ de la condition initiale y_{k-1} entraîne une perturbation proportionnelle de la solution dont le coefficient est minoré par const $|\mu_l(Q)|^{j+1-k}$. S'il existe un l tel que $|\mu_l| > 1$, le facteur $|\mu_l(Q)|^N = |\mu_l(Q)|^{X/h}$ correspondant au point $x_N = x_0 + X$ augmente brusquement avec la diminution de h. On obtient une solution très précise en réduisant l'erreur d'approximation, c'est-à-dire le pas d'intégration. Or, en vertu de ce qui précède on constate que lorsque le pas diminue, la méthode devient rapidement sensible aux perturbations des données initiales.

Ainsi, pour le schéma (6.7), $\max_i |\mu_i| = 5$. Supposons qu'en déterminant la valeur y_1 nous avons commis une erreur $\delta = 2^{-40}$. Alors dès $N = 18$, $|\mu_l(Q)|^N \delta$ est d'ordre 1 et l'erreur résultante est comparable à la solution même.

Il est également à craindre que pour μ_l p-uples, $p > 1$, la solution approchée sur le contour du cercle unité ne soit affectée d'une

erreur proportionnelle à $\delta N^{p-1} = \delta (Xh^{-1})^{p-1}$ qui est très grande pour h petits. Cette influence de l'erreur de calcul n'est pas une raison pour rejeter ces schémas étant donné que pour δ et N réels une erreur en δN^{p-1} est admissible. Montrons que ces méthodes sont inacceptables.

Soit μ_1 une racine p-uple de l'équation (4) et telle que $|\mu_1| = 1$.

Développons le premier membre de (3) en série de Taylor suivant les puissances de $\mu - \mu_1$ et Q; puisque $F(Q, \mu)$ est linéaire en Q, ce développement s'écrit

$$a_{00} + \sum_{j=1}^{\infty} a_{0j} (\mu - \mu_1)^j + \sum_{j=0}^{\infty} a_{1j} Q (\mu - \mu_1)^j = 0, \tag{5}$$

où

$$a_{ij} = F_Q i_{\mu} j (0, \mu_1)/j!.$$

Si μ_1 est une racine p-uple de (4), alors $a_{00} = \ldots = a_{0, p-1} = 0$. Compte tenu de cette circonstance, la relation (5) se met sous la forme

$$a_{0p} (\mu - \mu_1)^p + a_{10} Q + \psi (Q, \mu) = 0,$$

avec $\psi (Q, \mu)$ formée de termes de la forme const $(\mu - \mu_1)^{p+l}$ ou const $(\mu - \mu_1)^l Q$, > 0. Admettons que $a_{10} \neq 0$; la dernière égalité s'écrit alors

$$(\mu - \mu_1)^p = AQ + \varphi (Q, \mu), \tag{6}$$

$$A = -\frac{a_{10}}{a_{0p}} \neq 0, \infty,$$

avec $\varphi (Q, \mu)$ composée de termes de la même forme que ceux de $\varphi (Q, \mu)$, i.e.

$$\varphi (Q, \mu) = O((|Q| + |\mu - \mu_1|^p)\,|\mu - \mu_1|).$$

La théorie des fonctions d'une variable complexe apprend que l'équation (6) possède p racines $\mu_1 (Q), \ldots, \mu_p (Q)$

$$\mu_j (Q) = \mu_1 + (\sqrt[p]{AQ})_0 \exp\left(\frac{2\pi i j}{p}\right) + O(\sqrt[p]{|Q|^2}), \tag{7}$$

$(\sqrt[p]{AQ})_0$ étant une branche fixe de la fonction $\sqrt[p]{AQ}$.

Cette proposition se démontre comme suit. Considérons l'équation

$$\mu - \mu_1 = \varphi_j (Q, \mu), \tag{8}$$

avec $\varphi_j (Q, \mu)$ une branche de la fonction $\sqrt[p]{AQ + \varphi (Q, \mu)}$ qui se confond, lorsque $\mu = \mu_1$, avec $(\sqrt[p]{AQ})_0 \exp(2\pi i j/p)$. Prenons un $a > \sqrt[p]{A}$ quelconque et considérons le domaine $|\mu - \mu_1| \leqslant a \sqrt[p]{|Q|}$; on montre que pour $|Q| \leqslant C_2$, $C_2 > 0$ ce domaine vérifie l'estimation

$$\left|\frac{\partial \varphi_j}{\partial \mu}\right| \leqslant C_1 |Q|^{1/p} \leqslant 0{,}5.$$

Considérons le processus itératif de résolution de (8)

$$\mu^{(m+1)} = \mu_1 + \varphi_j (Q, \mu^{(m)}),$$

avec, au départ,

$$\mu^{(0)} = \mu_1 + (\sqrt[p]{AQ})_0 \exp\left(\frac{2\pi i j}{p}\right);$$

montrons que pour $|Q| \leqslant C_3$, $C_3 > 0$, toutes les itérées $\mu^{(m)}$ sont situées dans le domaine $|\mu^{(m)} - \mu_1| \leqslant a\sqrt[p]{|Q|}$, puis que, quel que soit m, on a

$$|\mu^{(m)} = \mu^{(0)}| \leqslant a_1 \sqrt[p]{|Q|^2}.$$

La convergence du processus itératif sous l'hypothèse

$$\left|\frac{\partial \varphi_j}{\partial \mu}\right| \leqslant 0{,}5$$

permet de dire que

$$\mu_j(Q) = \lim_{m \to \infty} \mu^{(m)}$$

satisfait à (8) et l'inégalité

$$|\mu_j(Q) - \mu^{(0)}| \leqslant a_1 \sqrt[p]{|Q|^2}.$$

D'où la validité de la relation (7).

Puisque le module d'un nombre complexe ne change pas s'il est multiplié par un nombre de module 1, on a en vertu de (7)

$$|\mu_j(Q)| = |\mu_j(Q)\mu_1^{-1}| = \left|(1 + \sqrt[p]{AQ})_0 \mu_1^{-1} \exp\left(\frac{2\pi i j}{p}\right) + \varepsilon_j(Q)\right|,$$

où

$$\varepsilon_j(Q) = O(\sqrt[p]{|Q|^2}).$$

Par ailleurs, le module d'un nombre complexe est au moins égal à sa partie réelle. C'est pourquoi

$$|\mu_j(Q)| \geqslant \operatorname{Re}\left(1 + (\sqrt[p]{AQ})_0 \mu_1^{-1} \exp\left(\frac{2\pi i j}{p}\right) + \varepsilon_j(Q)\right) =$$

$$= 1 + \operatorname{Re}\left((\sqrt[p]{AQ})_0 \mu_1^{-1} \exp\frac{2\pi i j}{p}\right) + \operatorname{Re}\varepsilon_j(Q).$$

En prenant au lieu de $y' = My$ l'équation $y' = -My$, celle-ci vérifie de même l'inégalité

$$|\mu_j(Q)| \geqslant 1 + \operatorname{Re}\left((\sqrt[p]{-AQ})_0 \mu_1^{-1} \exp\left(\frac{2\pi i j}{p}\right)\right) + \operatorname{Re}\varepsilon_j'(Q).$$

Considérons $2p$ nombres complexes

$$(\sqrt[p]{\pm AQ})_0 \mu_1^{-1} \exp\left(\frac{2\pi i j}{p}\right), \quad j = 1, \ldots, p.$$

Les points du plan complexe correspondant à ces nombres se trouvent sur la circonférence $|z| = \sqrt[p]{|AQ|}$ qu'ils divisent en $2p$ arcs égaux (fig. 8.7.1.) dont chacun est intercepté par un angle au centre de π/p. Comme $p \geqslant 2$, cet angle est au plus $\pi/2$. Par conséquent, l'un de ces nombres au moins possède

un argument égal au plus à $\pi/4$ en valeur absolue. Si $z = \rho(\cos\varphi + i\sin\varphi)$ et $|\varphi| \leqslant \frac{\pi}{4}$, alors

$$\operatorname{Re} z = \rho\cos\varphi \geqslant \rho\cos\left(\frac{\pi}{4}\right) = \frac{\rho}{\sqrt{2}}.$$

Et pour l'un de ces nombres

$$\operatorname{Re}\left(\left(\sqrt[p]{\pm AQ}\right)_0 \mu_1^{-1} \exp\left(\frac{2\pi i j}{p}\right)\right) \geqslant \frac{\sqrt[p]{|AQ|}}{\sqrt{2}}.$$

On aura donc, soit pour $y' = My$, soit pour $y' = -My$,

$$\max_j |\mu_j(Q)| \geqslant 1 + \frac{\sqrt[p]{|AQ|}}{\sqrt{2}} + \varepsilon(Q)$$

avec

$$\varepsilon(Q) = O\left(\sqrt[p]{|Q|^2}\right).$$

En se rappelant que $Q = Mh$, il en résulte

$$\max_j |\mu_j(Q)|^N \geqslant \left(1 + \frac{\sqrt[p]{|AQ|}}{\sqrt{2}} + \varepsilon(Q)\right)^{X/h} =$$

$$= \exp\left(\left(\frac{\sqrt[p]{|AQ|}}{\sqrt{2}} + O\left(\sqrt[p]{|Q|^2}\right)\right)\frac{X}{h}\right) = \exp\left(\frac{\sqrt[p]{|AM|}}{\sqrt{2}}\,\frac{X}{h^{\frac{p-1}{p}}} + O\left(\frac{1}{h^{\frac{p-2}{p}}}\right)\right).$$

Avec le fractionnement du pas h la dernière quantité tend vers l'infini plus vite que toute puissance de h. Aussi, même au bout d'un nombre faible de pas, les

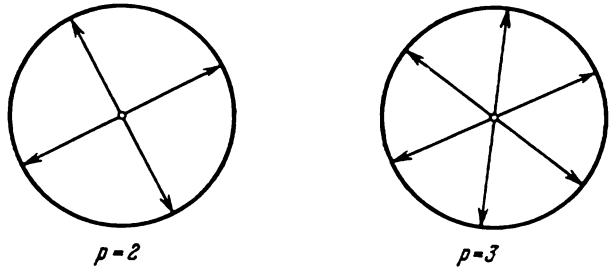


Fig. 8.7.1

perturbations par arrondi s'avèrent-elles comparables avec la solution cherchée, si bien qu'on renoncera également à des méthodes des différences finies dont l'équation caractéristique (4) admet des racines multiples de module unité.

Dans notre raisonnement nous avons admis $F_Q(0, \mu_1) \neq 0$. Dans le cas contraire, le schéma à $k+1$ valeurs correspond à un schéma à k valeurs. Ainsi, on obtient

$$y_j - 2y_{j-1} + y_{j-2} - h(f_{j-1} - f_{j-2}) = 0$$

en retranchant membre à membre les équations

$$y_j - y_{j-1} - hf_{j-1} = 0,$$
$$y_{j-1} - y_{j-2} - hf_{j-2} = 0,$$

correspondant au schéma de la méthode d'Euler.

A la lumière de l'exposé précédent on comprend l'importance de la *condition* α :

toutes les racines de l'équation caractéristique (4) *sont contenues dans le cercle unité et aucune racine multiple n'est située sur la circonférence de ce cercle.*

On pourrait penser que n'étaient-ce les erreurs d'arrondi et de données, la solution du problème aux différences finies convergerait vers celle du problème différentiel. Il n'en est rien : quel que soit le schéma aux différences ne vérifiant pas la condition α, on peut construire une équation différentielle possédant un second membre indéfiniment dérivable et telle que même en l'absence de ces erreurs la solution du problème aux différences finies soit éloignée de celle du problème différentiel.

Nous nous passerons de formulations et démonstrations rigoureuses des théorèmes correspondants. Car si une démonstration rigoureuse des propositions formulées sans équivoque a le don d'éclaircir le problème, elle cultive par contre un certain conservatisme dans l'approche du phénomène étudié. En énonçant par exemple qu'il existe, pour tout procédé ne satisfaisant pas à la condition α, une équation avec second membre régulier et telle qu'avec le fractionnement du pas du réseau la solution du problème discrétisé ne converge pas vers celle du problème différentiel dans le calcul sans arrondi, on suggère l'idée de renoncer une fois pour toutes à des procédés pareils. Une fois le théorème démontré, c'est cette conclusion qui est retenue. Est-elle juste? Actuellement, le calcul mathématique fait siennes des méthodes d'intégration négligeant la condition α et se servant, pour définir les valeurs y_j, d'une technique différant complètement de leur calcul par récurrence à l'aide d'un schéma aux différences. La résolution d'une classe de problèmes exigeait qu'on trouve la solution entachée d'une erreur relative d'environ 20 %, mais au bout d'un nombre de pas très faible. Les mathématiciens ont essayé en vain des méthodes vérifiant la condition α. Les ingénieurs ont alors proposé une technique négligeant cette condition mais donnant la solution avec la précision demandée pour un pas suffisamment important. Nous voyons que la cause de l'insuccès a été en l'occurrence une information redondante sur les méthodes de résolution. On a établi par la suite qu'une précision de 20 % ne vérifiait pas les nouvelles conditions imposées à la solution approchée. Pour améliorer la précision jusqu'à l'ordre de 1 % il a fallu utiliser tous les résultats théoriques et des méthodes vérifiant la condition α sur des machines à calculer plus puissantes.

De prime abord les schémas (1) de plus grand ordre d'approximation m semblent les plus rentables. Or, tous les schémas avec m grand ne satisfont pas à α.

THÉORÈMES. *Si le schéma*
est explicite, $m > k$,
implicite, k *impair*, $m > k + 1$,
ou implicite, k *pair*, $m > k + 2$,
il existe, parmi les racines de l'équation caractéristique (4), *une racine supérieure à* 1 *en module.*

§ 8. Evaluation d'erreur dans les méthodes des différences finies

Evaluons les erreurs sur les solutions approchées obtenues par des méthodes des différences finies du type (5.1) vérifiant α. Les approximations y_j des valeurs $y(x_j)$ sont liées non pas par la relation (6.1), mais par

$$\sum_{i=0}^{k} a_{-i} y_{j-i} - h \sum_{i=0}^{k} b_{-i} f(x_{j-i}, y_{j-i}) = \delta_j, \tag{1}$$

avec δ_j éventuellement non nul (l'origine de δ_j est éclaircie au § 4).

D'autre part, on a établi aux §§ 5, 6 que les valeurs $y(x_n)$ de la solution exacte du problème différentiel vérifient

$$\sum_{i=0}^{k} a_{-i} y(x_{j-i}) - h \sum_{i=0}^{k} b_{-i} f(x_{j-i}, y(x_{j-i})) = h r_j, \tag{2}$$

et on a vu que, pour un choix approprié des coefficients a_{-i} et b_{-i}, $r_j = O(h^m)$, $m > 0$. En retranchant (2) de (1) on trouve l'équation d'erreur $R_j = y_j - y(x_j)$. Selon la formule de Lagrange, on a

$$f(x_{j-i}, y_{j-i}) - f(x_{j-i}, y(x_{j-i})) = l_{j-i} R_{j-i}, \tag{3}$$

avec

$$l_{j-i} = f_y(x_{j-i}, \tilde{y}_{j-i}),$$

$\tilde{y}_{j-i}$ étant compris entre y_{j-i} et $y(x_{j-i})$. Compte tenu de (3), la différence de (1) et (2) s'écrit

$$\sum_{i=0}^{k} a_{-i} R_{j-i} - h \sum_{i=0}^{k} b_{-i} l_{j-i} R_{j-i} = g_j, \tag{4}$$

où $g_j = \delta_j - h r_j$.

THÉORÈME SUR L'ÉVALUATION D'ERREUR. *Supposons que le schéma aux différences vérifie la condition α et $|f_y| \leqslant L$ pour $x_0 \leqslant x \leqslant x_0 + X$. Dans la condition $x_0 \leqslant x_n \leqslant x_0 + X$, on a alors*

$$|R_n| \leqslant C(L, X)\Big(\max_{0 \leqslant i < k} |R_i| + \sum_{j=k}^{n} |g_j|\Big). \tag{5}$$

DÉMONSTRATION. Enonçons un cas particulier du lemme du § 3, chap. VI : supposons que toutes les valeurs propres de la matrice A se trouvent dans le cercle $|\lambda| \leqslant q$ et qu'il n'y a pas de racines multiples sur la circonférence ; il existe alors une matrice C telle que $\|D\|_1 \leqslant q$, où $D = C^{-1}AC$.

L'évaluation se fera plus aisément si l'on transforme (4) en une équation vectorielle à pas séparés. Pour fixer les idées, nous supposerons partout h tellement petit que $|hb_0L| \leqslant \left|\frac{a_0}{2}\right|$; le coefficient $a_0 - hb_0l_j$ affectant R_j dans (4) est alors $\left|\frac{a_0}{2}\right|$ au moins en module. En reportant à droite dans (4) tous les termes sans R_j et en divisant par le coefficient de R_j, il vient

$$R_j = \sum_{i=1}^{k} \frac{-a_{-i} + hb_{-i}l_{j-i}}{a_0 - hb_0l_j} R_{j-i} + \frac{g_j}{a_0 - hb_0l_j}. \tag{6}$$

Posons

$$\frac{-a_{-i} + hb_{-i}l_{j-i}}{a_0 - hb_0l_j} - \left(\frac{-a_{-i}}{a_0}\right) = hv_{ij};$$

évidemment,

$$v_{ij} = \frac{b_{-i}a_0l_{j-i} - a_{-i}b_0l_j}{(a_0 - hb_0l_j)\,a_0}, \quad |v_{ij}| \leqslant 2\,\frac{|b_{-i}a_0| + |a_{-i}b_0|}{|a_0|^2} L.$$

Introduisons les vecteurs

$$\mathbf{Z}_j = (R_j, \ldots, R_{j-k+1})^T.$$

La relation (6) est équivalente à

$$\mathbf{Z}_j = A\mathbf{Z}_{j-1} + hV_j\mathbf{Z}_j + \mathbf{W}_j, \tag{7}$$

où

$$A = \begin{pmatrix} -\frac{a_{-1}}{a_0} & -\frac{a_{-2}}{a_0} & \ldots & -\frac{a_{1-k}}{a_0} & -\frac{a_{-k}}{a_0} \\ 1 & 0 & \ldots & 0 & 0 \\ 0 & 1 & \ldots & 0 & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & 1 & 0 \end{pmatrix},$$

$$V_j = \begin{pmatrix} v_{1j} \ldots v_{kj} \\ 0 \ldots 0 \\ \ldots\ldots \\ 0 \ldots 0 \end{pmatrix}, \qquad \mathbf{W}_j = \begin{pmatrix} \frac{g_j}{a_0 - hb_0l_j} \\ 0 \\ \ldots \\ 0 \end{pmatrix}.$$

En effet, en égalant les premières composantes des vecteurs à droite et à gauche dans (7), on aboutit justement à (6) et en égalant les autres composantes à

$$R_{j-i} = R_{j-i}, \quad i = 1, \ldots, k-1.$$

Calculons le polynôme caractéristique de la matrice A :

$$P(\lambda) = \det(A - \lambda E) =$$

$$= \det \begin{pmatrix} -\frac{a_{-1}}{a_0} - \lambda & -\frac{a_{-2}}{a_0} & \dots & -\frac{a_{1-k}}{a_0} & -\frac{a_{-k}}{a_0} \\ 1 & -\lambda & \dots & 0 & 0 \\ 0 & 1 & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 1 & -\lambda \end{pmatrix}.$$

A cette fin, il faut multiplier par λ la première colonne et ajouter ce produit à la deuxième, puis multiplier par λ la deuxième colonne et ajouter ce produit à la troisième, et ainsi de suite. On a finalement

$$P(\lambda) = \det \begin{pmatrix} p_1(\lambda) & p_2(\lambda) & \dots & p_{k-1}(\lambda) & p_k(\lambda) \\ 1 & 0 & \dots & 0 & 0 \\ 0 & 1 & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 1 & 0 \end{pmatrix}$$

avec

$$p_1(\lambda) = -\frac{a_{-1}}{a_0} - \lambda, \quad p_2(\lambda) = -\frac{a_{-2}}{a_0} + \lambda\left(-\frac{a_{-1}}{a_0} - \lambda\right), \dots$$

$$p_k(\lambda) = -\frac{a_{-k}}{a_0} + \lambda\left(-\frac{a_{1-k}}{a_0} + \lambda\left(-\frac{a_{2-k}}{a_0} + \dots\right.\right.$$

$$\left.\left. \dots + \lambda\left(-\frac{a_{-1}}{a_0} - \lambda\right)\dots\right)\right).$$

Développons le déterminant suivant les éléments de la dernière colonne, il vient

$$(-1)^k P(\lambda) = -p_k(\lambda) = \lambda^k + \frac{a_{-1}}{a_0}\lambda^{k-1} + \dots + \frac{a_{-k}}{a_0} =$$

$$= \frac{1}{a_0}(a_0\lambda^k + a_{-1}\lambda^{k-1} + \dots + a_{-k}).$$

L'équation caractéristique de la matrice A est donc proportionnelle à l'équation caractéristique (7.5) du schéma aux différences. En vertu de la condition α, toutes les racines de l'équation caractéristique de A se situent dans le cercle $|z| \leqslant 1$ et il n'y a pas de racines multiples sur la circonférence. En ce qui concerne la matrice A, la condition du lemme est remplie pour $q = 1$. Par conséquent, il existe une matrice C telle que $C^{-1}AC = D$ et $\|D\|_1 \leqslant 1$. Effectuons le changement de variable $Z_n = Cz_n$ dans (7). La multiplication à gauche par C^{-1} réduit (7) à la forme

$$z_j = Dz_{j-1} + hv_j z_{j-1} + w_j, \tag{8}$$

où

$$D = C^{-1}AC, \quad v_j = C^{-1}V_jC, \quad w_j = C^{-1}W_j.$$

Les seuls éléments non nuls de la matrice V_j appartiennent à la première ligne ; aussi

$$\| V_j \|_1 \leqslant \left(2 \sum_{i=1}^{k} \frac{| b_{-i} a_0 | + | a_{-i} b_0 |}{| a_0 |^2} \right) L = vL$$

et

$$\| v_j \|_1 \leqslant \| C^{-1} \|_1 \| V_j \|_1 \| C \|_1 \leqslant vL \| C^{-1} \|_1 \| C \|_1.$$

Evidemment

$$\| w_j \|_1 \leqslant \| C^{-1} \|_1 \| W_j \|_1 = \| C^{-1} \|_1 \left| \frac{g_j}{a_0 - h b_0 l_j} \right| \leqslant 2 \| C^{-1} \|_1 \frac{| g_j |}{| a_0 |}.$$

Après avoir évalué les normes des termes dans le second membre de (8), on peut écrire l'inégalité

$$\| z_j \|_1 \leqslant \beta | g_i | + (1 + \gamma Lh) \| z_{j-1} \|_1, \tag{9}$$

où

$$\beta = 2 \frac{\| C^{-1} \|_1}{| a_0 |}, \quad \gamma = v \| C^{-1} \|_1 \| C \|_1.$$

Pour évaluer $\| z_n \|_1$ par $\| z_{k-1} \|_1$ et $| g_j |$ écrivons les inégalités (9) avec $j = n, \ldots, k$; dans le second membre de l'estimation de $\| z_n \|_1$ à l'aide de $\| z_{n-1} \|_1$, évaluons $\| z_{n-1} \|_1$ en fonction de $\| z_{n-2} \|_1$, puis $\| z_{n-2} \|_1$ par $\| z_{n-3} \|_1$ dans le second membre de l'inégalité ainsi obtenue, et ainsi de suite. Finalement,

$$\| z_n \|_1 \leqslant \beta | g_n | + (1 + \gamma Lh) (\beta | g_{n-1} | + (1 + \gamma Lh) (\beta | g_{n-2} | + \ldots$$
$$\ldots + (1 + \gamma Lh) (\beta | g_k | + (1 + \gamma Lh) \| z_{k-1} \|_1) \ldots)),$$

ou, ce qui revient au même,

$$\| z_n \|_1 \leqslant \beta \sum_{j=k}^{n} (1 + \gamma Lh)^{n-j} | g_j | + (1 + \gamma Lh)^{n-k+1} \| z_{k-1} \|_1. \tag{10}$$

Simplifions cette estimation en la rendant simultanément moins précise. Pour $x_0 \leqslant x_j \leqslant x_n \leqslant x_0 + X$, on a $(n - j)\, h \leqslant X$; c'est pourquoi

$$(1 + \gamma Lh)^{n-j} \leqslant \exp (\gamma Lh\, (n - j)) \leqslant \exp (\gamma LX).$$

Il découle alors de (10)

$$\| z_n \|_1 \leqslant \exp (\gamma LX) \left(\beta \sum_{j=k}^{n} | g_j | + \| z_{k-1} \|_1 \right). \tag{11}$$

Les inégalités

$$| R_n | \leqslant \| Z_n \|_1 \leqslant \| C \|_1 \| z_n \|_1,$$
$$\| z_{k-1} \|_1 \leqslant \| C^{-1} \|_1 \| Z_{k-1} \|_1 = \| C^{-1} \|_1 \max_{0 \leqslant i < k} | R_i |$$

sont justes, et, partant,

$$\| Z_n \|_1 \leqslant \| C \|_1 \exp(\gamma L X) \left(\beta \sum_{j=k}^{n} |g_j| + \| C^{-1} \|_1 \| Z_{k-1} \|_1\right). \quad (12)$$

On obtient maintenant l'estimation d'erreur

$$|R_n| \leqslant \exp(\gamma L X) \left(M_1 \sum_{j=k}^{n} |g_j| + M_2 \max_{0 \leqslant i < k} |R_i|\right),$$

où $M_1 = \beta \| C \|_1$, $M_2 = \| C \|_1 \| C^{-1} \|_1$ et γ sont des constantes ne dépendant que des coefficients a_{-l}, b_{-l} du schéma aux différences initial. En particulier, M_1 et M_2 dépendent des a_{-l} seuls. Le théorème est démontré.

Portant dans (12) la valeur $g_j = \delta_j - h r_j$, on trouve une estimation complète de l'erreur

$$|R_n| \leqslant \exp(\gamma L X) \left(M_1 \sum_{j=k}^{n} (|\delta_j| + h |r_j|) + M_2 \max_{0 \leqslant i < k} |R_i|\right). \quad (13)$$

Il en résulte que pour que la solution de l'équation aux différences converge vers celle de l'équation différentielle, il suffit que

$$\sum_{j=k}^{n} |\delta_j| \to 0, \quad h \sum_{j=k}^{n} |r_j| \to 0, \quad \max_{0 \leqslant i < k} |R_i| \to 0.$$

§ 9. Partie principale de l'erreur

Nous avons évalué les erreurs dans tous les schémas aux différences finies satisfaisant à α. Les estimations étant cependant trop fortes, la question de la méthode à choisir reste ouverte. Les schémas aux différences se prêtent à une caractérisation comparative plus détaillée si l'on met en évidence la partie principale de l'erreur sur la solution approchée.

Supposons que l'ordre d'approximation est strictement égal à $m \geqslant 1$ et que

1) le schéma aux différences vérifie la condition α et est normé, i. e. $b_0 + \ldots + b_{-k} = 1$;

2) le second membre $f(x, y)$ est suffisamment régulier, à savoir

$$|f_{x^{l_1} y^{l_2}}^{l_1+l_2}(x, y)| \leqslant L_{l_1 l_2} < \infty$$

dans le domaine G: $x_0 \leqslant x \leqslant x_0 + X$ quels que soient l_1, l_2, $0 \leqslant l_1 + l_2 \leqslant 2m$;

3) l'erreur sur les conditions initiales est d'ordre infinitésimal plus élevé que l'erreur d'approximation r_j, i. e.

$$|R_j| \leqslant \eta(h) h^m \quad \text{pour} \quad j = 0, \ldots, k-1,$$

$\eta(h)$ tendant vers zéro avec h;

4) les erreurs d'arrondi sont d'ordre infinitésimal plus élevé que les quantités $\rho_j = hr_j$ appelées *erreurs sur un pas*, c'est-à-dire que

$$|\delta_j| \leqslant \eta(h)\, h^{m+1} \quad \text{pour} \quad j = k, \ldots, N.$$

La dérivée $s+1$-ième de la solution s'exprime moyennant le second membre de l'équation et ses dérivées jusqu'à l'ordre s. Aussi a-t-on dans la condition 2)

$$|y^{(p)}(x)| \leqslant M_p < \infty \text{ pour } x_0 \leqslant x \leqslant x_0 + X \text{ et } p = 0, \ldots, 2m+1.$$

Selon (6.3), on a alors

$$|r_j - \sum_{p=0}^{2m} E_p h^{p-1} y^{(p)}(x_j)| \leqslant D_{2m+1} M_{2m+1} h^{2m}.$$

Vu que l'ordre d'approximation est m, i.e. $E_0 = \ldots = E_m = 0$, il s'ensuit

5) $$r_j = E_{m+1} y^{(m+1)}(x_j)\, h^m + \omega_j,$$

où $|\omega_j| \leqslant Qh^{m+1}$ uniformément par rapport à j lorsque $x_0 \leqslant x_j \leqslant x_0 + X$.

Bien que 5) soit la conséquence de la condition 2), nous la formulerons, pour la commodité de l'exposé, comme une condition indépendante.

THÉORÈME DE LA PARTIE PRINCIPALE DE L'ERREUR. *Dans les conditions* 1)-5) *et lorsque* $x_0 \leqslant x_n \leqslant x_0 + X$, *on a*

$$y_n - y(x_n) = E_{m+1} z(x_n)\, h^m + \varepsilon(x_n), \tag{1}$$

où $|\varepsilon(x_n)| \leqslant C(\eta(h) + h)\, h^m$;

$$z(x) = \int_{x_0}^{x} (-y^{(m+1)}(t)) \exp\left(\int_t^x f_y(\tau, y(\tau))\, d\tau\right) dt$$

est la solution de

$$z' = f_y(x, y(x))\, z - y^{(m+1)}(x)$$

avec la condition initiale $z(x_0) = 0$.

L'idée majeure de la démonstration est la suivante: l'équation (8.4) relativement à l'erreur $R_n = y_n - y(x_n)$ est ramenée à une forme qui entraîne que $R_n h^{-m}$ est solution de l'équation aux différences qui approche l'équation différentielle par rapport à $z(x)$ du théorème.

DÉMONSTRATION. Sous l'hypothèse du théorème la relation (8.4) s'écrit

$$\sum_{i=0}^{k} a_{-i} R_{j-i} - h \sum_{i=0}^{k} b_{-i} l_{j-i} R_{j-i} = \delta_j - hr_j = \\ = -E_{m+1} y^{(m+1)}(x_j)\, h^{m+1} + q_j, \tag{2}$$

avec

$$q_j = \delta_j - (hr_j - E_{m+1} y^{(m+1)}(x_j) h^{m+1});$$

conformément aux conditions 4), 5),

$$|q_j| \leqslant \eta(h) h^{m+1} + Qh^{m+2}.$$

Si l'on substitue dans le second membre de (8.13) les estimations de R_i, $0 \leqslant i < k$, δ_j, r_j, données par le théorème, il vient

$$|R_j| \leqslant d_1 h^m \text{ lorsque } x_0 \leqslant x_j \leqslant x_0 + X. \tag{3}$$

Rappelons-nous que $l_j = f_y(x_j, \widetilde{y}_j)$ avec $\widetilde{y}_j$ compris entre $y(x_j)$ et $y_j = y(x_j) + R_j$ et appliquons la formule de Lagrange relativement à y; nous obtenons

$$l_j - f_y(x_j, y(x_j)) = f_y(x_j, \widetilde{y}_j) - f_y(x_j, y(x_j)) = f_{yy}(x_j, \widetilde{\widetilde{y}}_j)(\widetilde{y}_j - y(x_j)).$$

Il en découle

$$|l_j - f_y(x_j, y(x_j))| \leqslant |L_{02}|\,|R_j| \leqslant d_1 L_{02} h^m \tag{4}$$

lorsque $x_0 \leqslant x_j \leqslant x_0 + X$. Utilisons

$$\sum_{i=0}^{k} b_{-i} = 1$$

et mettons l'égalité (2) sous la forme

$$\sum_{i=0}^{k} a_{-i} R_{j-i} - h \sum_{i=0}^{k} b_{-i} (f_y(x_{j-i}, y(x_{j-i})) R_{j-i} - E_{m+1} y^{(m+1)}(x_{j-i}) h^m) = \alpha_j, \tag{5}$$

où

$$\alpha_j = q_j + h \sum_{i=0}^{k} b_{-i} (l_{j-i} - f_y(x_{j-i}, y(x_{j-i}))) R_{j-i} + hE_{m+1} \Big(\sum_{i=0}^{k} b_{-i} (y^{(m+1)}(x_{j-i}) - y^{(m+1)}(x_j))\Big) h^m.$$

Conformément aux estimations (3), (4),

$$|(l_{j-i} - f_y(x_{j-i}, y(x_{j-i}))) R_{j-i}| \leqslant d_2 h^{2m}.$$

On a

$$|y^{(m+1)}(x_{j-i}) - y^{(m+1)}(x_j)| \leqslant M_{m+1} ih;$$

les estimations ci-dessus impliquent

$$\max_{x_0 \leqslant x_j \leqslant x_0 + X} |\alpha_j| = O(\eta(h) h^{m+1}) + O(h^{m+2}) + O(h^{2m+1}) = O((\eta(h) + h) h^{m+1}).$$

Divisons par h^{m+1} les deux membres de (5) et notons $v_j = R_j h^{-m}$, il vient

$$\sum_{i=0}^{k} a_{-i} v_{j-i} - h \sum_{i=0}^{k} b_{-i} (f_y (x_{j-i}, y(x_{j-i})) v_{j-i} - \\ - E_m y^{(m+1)} (x_{j-i})) = \beta_j, \qquad (6)$$

$$\beta_j = \alpha_j h^{-m}, \quad |\beta_j| \leqslant d_3 (\eta(h) + h) h.$$

Comparons les valeurs v_n avec la solution de l'équation différentielle

$$z' = f_y (x, y(x)) z - E_m y^{(m+1)} (x) \qquad (7)$$

dans la condition initiale $z(x_0) = 0$; nous avons

$$z(x) = -\int_{x_0}^{x} E_m y^{(m+1)} (t) \exp \left(\int_{t}^{x} f_y (\tau, y(\tau)) d\tau \right) dt. \qquad (8)$$

La relation (6) signifie que la fonction discrète v_j est une solution approchée de l'équation aux différences finies correspondant à cette équation différentielle linéaire. Servons nous de l'estimation (8.13) et substituons β_j à δ_j :

$$|v_n - z(x_n)| \leqslant \exp(\gamma L X) \left(M_1 \left(\sum_{j=k}^{n} |\beta_j| + h |\tilde{r}_j| \right) + \right. \\ \left. + M_2 \max_{0 \leqslant i < k} |v_i - z(x_i)| \right), \qquad (9)$$

$\tilde{r}_j$ étant l'erreur d'approximation

$$\tilde{r}_j = \sum_{i=0}^{k} \frac{a_{-i} z(x_{j-i})}{h} - \sum_{i=0}^{k} b_{-i} z'(x_{j-i}),$$

qui correspond à l'équation

$$z' = f_y (x, y(x)) z - y^{(m+1)} (x).$$

La solution $y(x)$ est $(2m+1)$ fois dérivable par rapport à x ; il en découle que le second membre de (7) l'est m fois. La fonction $z(x)$ est alors $m+1$ fois dérivable. En vertu des développements du § 5, $|\tilde{r}_j| \leqslant d_4 h^m$ lorsque $x_0 \leqslant x_j \leqslant x_0 + X$. D'où

$$\sum_{j=k}^{n} h |\tilde{r}_j| \leqslant X \max_j |\tilde{r}_j| = XO(h^m).$$

Comme $|\beta_j| \leqslant d_3 (\eta(h) + h) h$, on a

$$\sum_{j=1}^{n} |\beta_j| \leqslant d_3 (\eta(h) + h) nh \leqslant d_3 X (\eta(h) + h)$$

lorspue $x_0 \leqslant x_n \leqslant x_0 + X$.

D'après l'hypothèse du théorème, $v_i = R_i h^{-m} = O(\eta(h))$ pour $0 \leqslant i < k$; la forme explicite de $z(x)$ entraîne également $z(x_i) = O(h)$ pour $0 \leqslant i < k$, et donc

$$\max_{0 \leqslant i < k} | v_i - z(x_i) | = O(\eta(h) + h).$$

Utilisons ces inégalités pour évaluer le second membre de (9); nous obtenons

$$\max_n | v_n - z(x_n) | = O(\eta(h) + h).$$

La multiplication par h^m fournit l'affirmation cherchée du théorème.

Le théorème de la partie principale de l'erreur autorise formellement à appliquer la règle d'évaluation pratique de l'erreur de Runge (voir §§ 13, 15, chap. III) pour h et les erreurs d'arrondi δ_j suffisamment petits. Dans le cas considéré, il importe cependant d'être plus prudent que lorsqu'il s'agissait de calculer des intégrales.

En effet, la relation

$$| E_{m+1} z(x) h^m | \gg | \varepsilon(x) |, \tag{10}$$

qui est nécessaire pour que $E_{m+1} z(x) h^m$ soit effectivement la partie principale de l'erreur, n'a parfois lieu que pour h très faibles et dans la condition invraisemblable de δ_j petit.

Ce que nous venons de dire n'est vrai que pour certaines méthodes. Dans les méthodes à pas séparés, on constate assez souvent une prédominance réelle de la partie principale de l'erreur (4.12) par rapport aux termes d'ordre supérieur. Parmi les procédés utilisant les différences finies indiquons ceux dans lesquels une racine de l'équation caractéristique (7.4) vaut 1 et les autres sont strictement intérieures au cercle unité (les méthodes d'Adams en font partie). Pour les méthodes d'Adams et de Runge-Kutta la relation (10) est vérifiée dès que les pas h sont relativement acceptables. Par ailleurs, des calculs réels selon ces méthodes se font assez souvent avec des pas grands qui ne satisfont pas toujours à (10).

En l'absence de condition (10), on peut dire que l'erreur possède réellement une partie principale qui ne coïncide cependant pas avec $E_{m+1} z(x) h^m$. Par exemple, la partie principale effective pour le schéma $\frac{y_j - y_{j-2}}{2h} - f(x_{j-1}, y_{j-1}) = 0$ peut être une certaine composante $\varepsilon(x)$ de l'ordre de h^{m+1}. Le paragraphe suivant confirmera dans une certaine mesure nos raisonnements.

On peut dire, pour conclure, que la comparaison des résultats des calculs effectués avec des pas différents est le procédé principal (et d'emploi relativement fréquent) qui permet de se faire une idée de l'erreur réelle.

L'analyse faite aux §§ 8, 9 semble impliquer que tous les schémas d'un même ordre d'approximation et tels que la condition α soit satisfaite sont équivalents quant à leur mise en œuvre pratique: selon les constructions du § 8, les conditions de convergence des solutions approchées ont un aspect sensiblement identique et, d'après le théorème que nous venons de démontrer, les parties principales de l'erreur sont proportionnelles.

Cette analyse s'avère néanmoins insuffisante ; nous avons examiné le comportement asymptotique de l'erreur avec un segment d'intégration $[x_0, x_0 + X]$ fixe et un pas du réseau tendant vers zéro. Or, de nombreux problèmes se caractérisent par un intervalle d'intégration très grand si bien qu'on a intérêt à connaître non seulement l'allure de l'erreur pour un pas qui va en diminuant, mais aussi pour un segment d'intégration important. Pour éclaircir cette question, examinons une fois de plus des problèmes modèles.

§ 10. Etude des propriétés des méthodes des différences finies sur des modèles plus précis

Comparons deux schémas aux différences finies : la *méthode d'Euler*

$$\frac{y_j - y_{j-1}}{h} - f(x_{j-1}, y_{j-1}) = 0 \tag{1}$$

et une variante de la *méthode de Milne*

$$\frac{y_j - y_{j-2}}{2h} - f(x_{j-1}, y_{j-1}) = 0. \tag{2}$$

Comme au § 5, substituons à y_j les valeurs $y(x_j)$ de la solution exacte et obtenons, après avoir développé le premier membre en série de Taylor suivant les puissances de h, l'erreur d'approximation

$$r_j^{(1)} = \frac{y(x_j) - y(x_{j-1})}{h} - f(x_{j-1}, y(x_{j-1})) = y''(x_j)\frac{h}{2} + O(h^2), \tag{3}$$

$$r_j^{(2)} = \frac{y(x_j) - y(x_{j-2})}{2h} - f(x_{j-1}, y(x_{j-1})) = y'''\frac{(x_j)}{6}h^2 + O(h^3), \tag{4}$$

pour le premier et le deuxième schéma respectivement. Le calcul d'après le premier schéma exige qu'on connaisse au départ une valeur approchée de $y(x_0)$, et dans le cas du deuxième s'y ajoute la connaissance de la valeur approchée de $y(x_1)$. Supposons ces valeurs données sans erreur et le calcul même se passant d'arrondi. Le premier schéma a pour équation caractéristique $\mu - 1 = 0$ et le deuxième $\mu^2 - 1 = 0$. Les deux équations vérifient la condition α. Supposons que le second membre $f(x, y)$ satisfait à toutes les hypothèses du théorème sur la partie principale de l'erreur (§ 9) ; on a alors, conformément au théorème,

$$y_n^1 - y(x_n) = -\frac{h}{2}\int_{x_0}^{x_n} y''(t) \exp\left(\int_t^{x_n} f_y(\tau, y(\tau))d\tau\right) dt \tag{5}$$

pour le premier schéma, et

$$y_n^2 - y(x_n) = -\frac{h^2}{6}\int_{x_0}^{x_n} y'''(t)\exp\left(\int_t^{x_n} f_y(\tau, y(\tau))\,d\tau\right)dt \qquad (6)$$

pour le deuxième.

Proposons-nous maintenant d'intégrer le modèle $y' = My$ avec la condition initiale $y(0) = 1$. Son second membre vérifie toutes les hypothèses du théorème du paragraphe précédent; évidemment $f_y = M$, $y^{(k)}(x) = M^{(k)}\exp(Mx)$. Les relations (5) et (6) donnent respectivement

$$y_n^1 - y(x_n) = -\frac{h}{2}\int_0^{x_n} M^2\exp(Mt)\exp\left(\int_t^{x_n} M\,d\tau\right)dt =$$

$$= -\frac{M^2 x_n h}{2}\exp(Mx_n) + O(h^2),$$

$$y_n^2 - y(x_n) = -\frac{M^3 x_n h^2}{6}\exp(Mx_n) + O(h^3).$$

La partie principale de l'erreur dans le deuxième schéma diffère de celle dans le premier par le facteur $Mh/3$. Une comparaison selon le critère de partie principale révélerait l'avantage du deuxième schéma lorsque $|Mh/3| < 1$.

Cette analyse s'avère néanmoins trop sommaire. Les solutions des équations aux différences finies (1), (2) peuvent s'expliciter. On a dans le cas du premier schéma

$$\frac{y_n^1 - y_{n-1}^1}{h} - My_{n-1}^1 = 0,$$

alors

$$y_n^1 = (1 + Mh)\,y_{n-1}^1,$$

et donc

$$y_n^1 = (1 + Mh)^n.$$

Dans le deuxième cas, on a l'équation

$$\frac{y_n^2 - y_{n-2}^2}{2h} - My_{n-1}^2 = 0$$

dont la solution générale s'écrit

$$y_n^2 = C_1\mu_1^n + C_2\mu_2^n,$$

avec $\mu_{1,2} = Mh \pm \sqrt{1 + M^2h^2}$ racines de

$$\mu^2 - 2Mh\mu - 1 = 0.$$

Les conditions de coïncidence en $x = x_0 = 0$ et $x = x_1 = h$ avec la solution exacte donnent un système d'équations en C_1 et C_2:

$$C_1 + C_2 = 1, \quad C_1\mu_1 + C_2\mu_2 = \exp(Mh);$$

d'où

$$C_1=\frac{\exp(Mh)-\mu_2}{\mu_1-\mu_2},\quad C_2=\frac{\exp(Mh)-\mu_1}{\mu_2-\mu_1};$$

les racines μ_1, μ_2, et, partant, les coefficients C_1, C_2 sont fonctions de $Mh=Q$; développons-les suivant les puissances de Q; il vient

$$\sqrt{1+M^2h^2}=1+\frac{Q^2}{2}+O(Q^4),$$

$$\mu_1=1+Q+\frac{Q^2}{2}+O(Q^4)=\exp Q-\frac{Q^3}{6}+O(Q^4)=$$

$$=\exp\left(Q-\frac{Q^3}{6}+O(Q^4)\right),$$

$$\mu_2=-1+Q-\frac{Q^2}{2}+O(Q^4)=-\exp(-Q)-\frac{Q^3}{6}+O(Q^4)=$$

$$=-\exp\left(-Q+\frac{Q^3}{6}+O(Q^4)\right), \quad (7)$$

$$\mu_1-\mu_2=2+Q^2+O(Q^4),$$

$$C_1=\frac{1+Q+\frac{Q^2}{2}+\frac{Q^3}{6}+1-Q+\frac{Q^2}{2}+O(Q^4)}{2+Q^2+O(Q^4)}=$$

$$=\frac{2+Q^2+\frac{Q^3}{6}+O(Q^4)}{2+Q^2+O(Q^4)}=1+\frac{Q^3}{12}+O(Q^4),$$

$$C_2=1-C_1=-\frac{Q^3}{12}+O(Q^4).$$

Donc,

$$y_n^2=\left(1+\frac{Q^3}{12}+O(Q^4)\right)\left(\exp\left(Q-\frac{Q^3}{6}+O(Q^4)\right)\right)^n-$$

$$-\left(\frac{Q^3}{12}+O(Q^4)\right)(-1)^n\left(\exp\left(-Q+\frac{Q^3}{6}+O(Q^4)\right)\right)^n;$$

on suppose ensuite $|Q|\leqslant$ const. On a l'égalité évidente

$$1+Q=\exp\left(Q-\frac{Q^2}{2}+O(Q^3)\right).$$

Aussi le rapport de la solution approchée y_n^1 à la solution exacte $y(x_n)=\exp(Mx_n)=\exp(Qn)$ du premier schéma s'écrit

$$\frac{y_n^1}{y(x_n)}=\frac{\exp\left(\left(Q-\frac{Q^2}{2}+O(Q^3)\right)n\right)}{\exp(Qn)}=\exp\left(\left(-\frac{Q^2}{2}+O(Q^3)\right)n\right)=$$

$$=1-\frac{Q^2n}{2}+O(Q^3n)+O(Q^4n^2)=$$

$$=1-\frac{M^2x_n}{2}h+O(M^3x_n(1+|M|x_n)h^2). \quad (8)$$

Dans le deuxième cas

$$\frac{y_n^2}{y(x_n)} = \left(1+\frac{Q^3}{12}+O(Q^4)\right)\exp\left(\left(-\frac{Q^3}{6}+O(Q^4)\right)n\right) - \\ -\left(\frac{Q^3}{12}+O(Q^4)\right)\exp\left(\left(-2Q+\frac{Q^3}{6}+O(Q^4)\right)n\right) = \\ = 1+O(M^3x_nh^2)-\frac{(Mh)^3}{12}(1+O(Mh))\exp(-2Mx_n(1+O(Mh)^2)). \tag{9}$$

Si $M \geqslant 0$, il en résulte

$$\frac{y_n^2}{y(x_n)} = 1+O(M^3x_nh^2); \tag{10}$$

pour Mh petits, l'estimation (10) est meilleure que (8) quelle que soit la longueur du segment d'intégration. Si $M < 0$, le deuxième terme à droite dans (9) renferme un facteur qui croît brusquement avec x_n. Par exemple, pour $M = -1$, $h = 0{,}01$, $x_n = 10$, on a

$$\left|\frac{y_n^2}{y(x_n)}\right| > 30.$$

Tandis qu'on a, lorsqu'il s'agit de la formule de calcul (1),

$$\frac{y_n^1}{y(x_n)} = 0{,}95\ldots$$

Pourquoi la formule (1) est-elle meilleure pour $M < 0$, alors qu'elle n'est exacte qu'à l'ordre 1 ?

Les solutions de l'équation différentielle considérée ont la forme :

$$\text{const}\cdot\exp(Mx),$$

et elles se multiplient par $\exp(Mh) = \exp Q$ lorsqu'on passe d'un point x_n au point suivant. L'équation aux différences finies correspondant au schéma (2) admet des solutions particulières de deux formes $C_1\mu_1^n$ et $C_2\mu_2^n$. Selon (7),

$$\mu_1 = \exp(Mh) + O(|Mh|^3), \tag{11}$$

et la composante de la solution correspondant à cette racine est multipliée à chaque pas par une quantité voisine de $\exp(Mh)$. La composante associée à la seconde racine se multiplie à chaque pas par μ_2 qui est sensiblement différent de $\exp(Mh)$ (on peut dire que c'est une composante parasite). Si $M > 0$, alors $|\mu_2| < |\mu_1|$ et l'influence de la composante parasite décroît à chaque pas. Si par contre $M < 0$, alors $|\mu_2| > |\mu_1|$ et l'influence de la composante augmente rapidement et perturbe catastrophiquement le résultat.

Dans la pratique, on aborde non pas une équation isolée, mais tout un système. Si l'écart de deux solutions quelconques du système

croît, l'équation $y' = My$, $M > 0$, est acceptée comme modèle sur lequel on essaye les méthodes de résolution du système. Si deux solutions quelconques se rapprochent dans le temps, le modèle est souvent $y' = My$, $M < 0$. Les solutions d'équations différentielles ne manifestent pas fréquemment de tendance bien prononcée à diverger ou converger à chaque instant, cette tendance se révélant globalement. On exige donc que la méthode en question intègre suffisamment bien l'équation $y' = My$ que M soit supérieur ou inférieur à zéro. Le cas $M < 0$ est astreint à des conditions particulièrement importantes vu que le rapprochement caractéristique des solutions est spécial aux problèmes dissipatifs qui constituent une grande fraction des problèmes d'applications. Il existe suffisamment de problèmes dont les solutions convergent ou divergent relativement lentement (cette classe comprend des problèmes sans dissipation d'énergie).

L'équation $y' = My$, $M < 0$ ou $M > 0$, ne peut servir de modèle pour essayer les méthodes de résolution de ces problèmes en raison du comportement différent de leurs solutions.

On prend en l'occurrence pour modèle l'intégration de $y' = My$ par rapport à la fonction y à valeurs complexes pour M imaginaire pur : $M = is$, s réel. Les solutions $y(x) = y(0) \exp(isx)$ de cette équation et leurs différences restent de module constant.

Dans le cas du schéma (1)

$$y_n = (1 + ish)\, y_{n-1}, \quad |1 + ish| = \sqrt{1 + s^2h^2} > 1,$$

par conséquent,

$$|y_n| = \sqrt{1 + s^2h^2}\, |y_{n-1}|$$

et les solutions de l'équation aux différences augmentent en module.

Dans le cas du schéma (2) les quantités y_n vérifient l'équation aux différences

$$y_n - y_{n-2} - 2shiy_{n-1} = 0,$$

et l'équation caractéristique correspondante

$$\mu^2 - 2shi\mu - 1 = 0$$

admet pour racines

$$shi \pm \sqrt{1 - s^2h^2}$$

de module 1 ; par conséquent, la solution de l'équation aux différences oscille sans présenter de tendance stable à augmenter ou à décroître. Ce comportement de la solution est préférable à celui de la solution du schéma (1) sans parler de l'ordre de précision plus élevé du schéma (2). L'allure de la solution du problème discrétisé peut sensiblement changer lors du passage au cas $M = M(x)$. S'il s'agit d'équations non linéaires, la situation s'embrouille plus

encore. Aussi la mise au point des méthodes de résolution des problèmes avec conservation d'énergie sur les seuls modèles du type $y' = My$ s'avère insuffisante. Sur des intervalles de temps considérables on intègre parfois des équations différentielles correspondant aux problèmes avec conservation d'énergie à l'aide de schémas spéciaux tels que des analogues discrets des intégrales de l'énergie pris sur leurs solutions restent constants. Pour choisir les méthodes à incorporer dans des programmes standard on se borne généralement à une analyse du comportement des solutions des équations aux différences correspondant à $y' = My$ avec $M > 0$ ou < 0.

La solution de l'équation différentielle $y' = My$ est multipliée à chaque pas par $\exp(Mh)$. Dans les exemples examinés plus haut, l'équation caractéristique de l'équation aux différences correspondante avait toujours une racine voisine de $\exp(Mh)$; c'est précisément cette racine qui rapproche les solutions de l'équation différentielle et celles aux différences.

THÉORÈME. *Pour que le schéma aux différences finies normé*

$$\sum_{i=0}^{k} \frac{a_{-i} y_{n-i}}{h} - \sum_{i=0}^{k} b_{-i} f(x_{n-i}, y_{n-i}) = 0 \tag{12}$$

soit exact à un ordre strictement égal à m, il faut et il suffit qu'il y ait parmi les racines de l'équation

$$\sum_{i=0}^{k} (a_{-i} - Mhb_{-i}) \mu^{k-i} = 0 \tag{13}$$

la racine

$$\mu = \exp(Mh) - E_{m+1} (Mh)^{m+1} + O((Mh)^{m+2}), \quad E_{m+1} \neq 0.$$

DÉMONSTRATION. Supposons que le schéma aux différences (12) soit approché à un ordre strictement égal à m, c'est-à-dire que pour une solution suffisamment régulière arbitraire

$$\sum_{i=0}^{k} \frac{a_{-i} y(x_{n-i})}{h} - \sum_{i=0}^{k} b_{-i} f(x_{n-i}, y(x_{n-i})) =$$

$$= E_{m+1} y^{(m+1)}(x_n) h^m + O(h^{m+1}), \; E_{m+1} \neq 0. \tag{14}$$

Posons $f(x, y) = My$ et portons dans (14) la solution $y(x) = \exp(Mx)$; lorsque $x_q \equiv qh$ et $n = k$, on trouve

$$\sum_{i=0}^{k} \frac{a_{-i} \exp(M(k-i)h)}{h} - \sum_{i=0}^{k} b_{-i} M \exp(M(k-i)h) =$$

$$= E_{m+1} M^{m+1} h^m \exp(Mkh) + O(h^{m+1})$$

Une multiplication par h donne

$$\sum_{i=0}^{k}(a_{-i}-Mhb_{-i})\exp(Mh(k-i))=E_{m+1}(Mh)^{m+1}\exp(Mkh)+O(h^{m+2})=$$

$$=E_{m+1}(Mh)^{m+1}+O(h^{m+2}).$$

Tous les termes explicites des deux membres sont fonctions de $Q=Mh$; c'est pourquoi la quantité $O(h^{m+2})$ dépend en fait de Mh, i.e. cette quantité doit s'écrire $O((Mh)^{m+2})$. La dernière relation se récrit donc

$$\sum_{i=0}^{k}(a_{-i}-Qb_{-i})(\exp Q)^{k-i}=E_{m+1}Q^{m+1}+O(Q^{m+2}),$$

ou encore

$$P(\exp Q)=E_{m+1}Q^{m+1}+r_{m+2},$$

où

$$P(\mu)=\sum_{i=0}^{k}(a_{-i}-b_{-i}Q)\mu^{k-i},\quad r_{m+2}=O(Q^{m+2}).$$

Considérons le segment $|\mu-1|\leqslant 2|Q|$ et évaluons sur lui la dérivée

$$P'(\mu)=\sum_{i=0}^{k}(a_{-i}-Qb_{-i})(k-i)\mu^{k-i-1}$$

Pour les points de $|\mu-1|\leqslant 2|Q|$

$$\mu^{k-i-1}=1+O(Q),$$

par conséquent,

$$P'(\mu)=\sum_{i=0}^{k}a_{-i}(k-i)+O(Q).$$

Etant données les conditions $E_0=E_1=0$ et vu que le schéma est normé,

$$\sum_{i=0}^{k}a_{-i}=0\quad\text{et}\quad\sum_{i=0}^{k}a_{-i}(-i)=1$$

et la dernière égalité se met sous la forme

$$P(\mu)=1+O(Q),$$

ou

$$1-C|Q|\leqslant P'(\mu)\leqslant 1+C|Q|\text{ pour }|\mu-1|\leqslant 2|Q|.$$

Posons

$$\mu^{1,2}=\exp Q-\frac{E_{m+1}Q^{m+1}+r_{m+2}}{1\mp CQ}.$$

Evidemment,

$$\mu^{1,2}=1+Q+O(Q^2).$$

Nous allons supposer $|Q|$ tellement petit que $|\exp Q-1|\leqslant 2|Q|$

$$|\mu^{1,2}-1|\leqslant 2|Q|,\quad C|Q|<1\quad\text{et}\quad|r_{m+2}|<|E_{m+1}Q^{m+1}|.$$

Pour fixer les idées, prenons le cas $E_{m+1}>0$, $Q>0$. Alors

$$\mu^1<\mu^2<\exp Q.$$

On a les inégalités évidentes

$$P(\exp Q)-P(\mu^2)\leqslant\Big(\max_{[\mu^2,\ \exp Q]}P'(\mu)\Big)(\exp Q-\mu^2)\leqslant$$

$$\leqslant(1+C|Q|)\frac{E_{m+1}Q^{m+1}+r_{m+2}}{1+C|Q|}=P(\exp Q),$$

$$P(\exp Q)-P(\mu^1)\geqslant\Big(\min_{[\mu^1,\ \exp Q]}P'(\mu)\Big)(\exp Q-\mu^1)\geqslant$$

$$\geqslant(1-C|Q|)\frac{E_{m+1}Q^{m+1}+r_{m+2}}{1-C|Q|}=P(\exp Q).$$

On en conclut que

$$P(\mu^1)\leqslant 0\leqslant P(\mu^2);$$

selon le théorème de Rolle, l'équation $P(\mu)=0$ a une racine dans les limites $\mu^1\leqslant\mu\leqslant\mu^2$. Les quantités μ^1, μ^2 s'écrivent par construction:

$$\exp Q-E_{m+1}Q^{m+1}+O(Q^{m+2}),$$

donc,

$$\mu=\exp Mh-E_{m+1}(Mh)^{m+1}+O((Mh)^{m+2}).$$

Le théorème est démontré dans un sens.

Supposons ensuite que $P(\mu)=0$ admet une racine

$$\mu=\exp Q-E_{m+1}Q^{m+1}+O(Q^{m+2}),\ E_{m+1}\neq 0.$$

Raisonnant par remontée (nous en laissons le soin au lecteur) on constate que

$$P(\exp Q)=E_{m+1}Q^{m+1}+O(Q^{m+2}).$$

Développons le premier membre en série de Taylor suivant les puissances de Q. Puisque le terme principal du second membre est $E_{m+1}Q^{m+1}$, les coefficients des puissances de Q^l s'annulent pour $0\leqslant l\leqslant m$ et le coefficient de Q^{m+1} vaut E_{m+1}. On trouve les relations

$$E_0=\sum_{i=0}^{k}a_{-i}=0,$$

$$E_l=\sum_{i=0}^{k}\frac{a_{-i}(-i)^l}{l!}-\sum_{i=0}^{k}\frac{b_{-i}(-i)^{l-1}}{(l-1)!}=0,\quad l=1,\ \ldots,\ m,$$

$$\sum_{i=0}^{k}\frac{a_{-i}(-i)^{m+1}}{(m+1)!}-\sum_{i=0}^{k}\frac{b_{-i}(-i)^m}{m!}=E_{m+1}\neq 0.$$

D'après les constructions du § 5, cela veut justement dire que les solutions régulières des équations différentielles vérifient (14). C.Q.F.D.

Si le schéma aux différences approxime l'équation différentielle, alors $m\geqslant 1$; donc, l'équation (13) possède toujours la racine $\mu=$ $=\exp(Mh)+O((Mh)^2)$. Lorsque $Mh\to 0$, cette racine tend vers

la racine $\mu = 1$ de l'équation caractéristique du schéma (7.4). Actuellement, on utilise des schémas aux différences dont l'équation caractéristique (7.4) a toutes ses autres racines nulles.

Cela signifie que $\sum_{i=0}^{k} a_{-i}\mu^{k-i} = (\mu - 1)\,\mu^{k-1}$, donc $a_0 = 1$, $a_{-1} = -1$. Pour de tels schémas les racines parasites de (13) tendent vers zéro avec Mh. La classe de ces méthodes comprend entre autres les formules d'Adams explicites et implicites que nous avons définies plus haut. Sur l'exemple des premières établissons plus scrupuleusement les valeurs Mh pour lesquelles les racines parasites du schéma aux différences exercent une influence peu importante.

Les formules d'Adams explicites s'écrivent

$$y_n - y_{n-1} - h\sum_{i=0}^{k-1} \gamma_i \nabla^i f_{n-1} = 0.$$

Substituant $f(x, y) = My$, on trouve l'équation aux différences

$$y_n - y_{n-1} - Mh\sum_{i=0}^{k-1} \gamma_i \nabla^i y_{n-1} = 0,$$

dont l'équation caractéristique

$$P(\mu) = \mu^k - \mu^{k-1} - Mh\sum_{i=0}^{k-1} \gamma_i (\mu - 1)^i \mu^{k-1-i} = 0.$$

Traitons le cas $M < 0$. On a

$$P(-1) = (-1)^k - (-1)^{k-1} - Mh\sum_{i=0}^{k-1} \gamma_i (-2)^i (-1^{k-1-i}) =$$

$$= (-1)^k [2 + Mh\sum_{i=0}^{k-1} \gamma_i 2^i].$$

Le polynôme $P(\mu)$ a le signe $(-1)^k$ pour des μ négatifs de module suffisamment grand. Si l'expression entre crochets est négative, la valeur $P(-1)$ est de signe contraire, donc $P(\mu)$ possède la racine $\mu < -1$. La solution de l'équation aux différences contient alors une composante qui s'amplifie alors que celle de l'équation différentielle décroît. Aussi pour obtenir de bons résultats en intégrant sur un intervalle de temps important est-il nécessaire que soit réalisée la condition

$$2 + Mh\sum_{i=0}^{k-1} \gamma_i 2^i > 0. \tag{15}$$

Nous avons montré que γ_i est de l'ordre de $\frac{1}{\ln i}$, par conséquent $\sum_{i=1}^{k-1} \gamma_i 2^i$ croît comme $\frac{2^k}{\ln k}$. Conclusion: la condition (15) imposée à Mh, $M < 0$, devient de plus en plus sévère avec l'augmentation de l'ordre d'approximation k.

Lorsqu'on intègre les solutions régulières par la méthode d'Adams avec une recherche automatique du pas, on observe une situation curieuse. Vu une régularité suffisante de la solution, le pas commence à augmenter. Lorsque $f_y < 0$, on constate une prédominance des racines parasites de (13) sur la racine principale. La fonction discrète obtenue perd, par conséquent, sa régularité et le pas se fractionne. Avec la diminution du pas, la fonction redevient régulière, puis le pas augmente, et ainsi de suite. Ce phénomène a été mis en évidence lors de la comparaison du temps de calcul réel par la méthode d'Adams avec sa valeur théorique: le temps observé a notablement dépassé le temps théorique par suite d'un volume important de calcul dû à l'augmentation et la fragmentation périodiques du pas.

§ 11. Intégration de systèmes d'équations

La plupart des constructions ci-dessus s'étendent telles quelles au cas des systèmes d'équations

$$\mathbf{y}' = \mathbf{f}(x, \mathbf{y}).$$

La seule différence, formelle d'ailleurs, est que dans les relations correspondantes figurent des tenseurs ou matrices au lieu de grandeurs scalaires. Ainsi, l'intégration d'une équation différentielle par la série de Taylor exigeait le calcul des dérivées

$$y'' = f_x + f_y y',$$
$$y''' = f_{xx} + 2f_{xy}y' + f_{yy}(y')^2 + f_y y'', \text{ etc.}$$

Soit

$$\mathbf{y} = (y_1, \ldots, y_q)^T, \quad \mathbf{f} = (f_1, \ldots, f_q)^T.$$

S'il s'agit d'un système d'équations différentielles, les dérivées $y_i^{(k)}$ s'obtiennent à partir des relations analogues, résultat de la dérivation du système initial

$$y_i'' = \frac{d}{dx}(f_i(x, \mathbf{y})) = \frac{\partial f_i}{\partial x} + \sum_{j_1=1}^{q} \frac{\partial f_i}{\partial y_{j_1}} \frac{dy_{j_1}}{dx},$$

$$y_i''' = \frac{d}{dx}(y_i'') = \frac{\partial^2 f_i}{\partial x^2} +$$

$$+ 2\sum_{j_1=1}^{q} \frac{\partial^2 f_i}{\partial x\, \partial y_{j_1}} \frac{dy_{j_1}}{dx} + \sum_{j_1=1}^{q}\sum_{j_2=1}^{q} \frac{\partial^2 f_i}{\partial y_{j_1}\, \partial y_{j_2}} \frac{dy_{j_1}}{dx}\frac{dy_{j_2}}{dx} +$$

$$+ \sum_{j_1=1}^{q} \frac{\partial f_i}{\partial y_{j_1}} \frac{d^2 y_{j_1}}{dx^2}, \ldots,$$

q étant l'ordre du système.

L'ensemble de ces relations s'écrit sous forme tensorielle

$$\mathbf{y}'' = \mathbf{f}_x + f_y \mathbf{y}',$$
$$\mathbf{y}''' = \mathbf{f}_{xx} + 2f_{xy}\mathbf{y}' + f_{yy}\mathbf{y}'\mathbf{y}' + f_y \mathbf{y}'', \ldots;$$

$f_{x^n y^m}$ est le tenseur

$$\frac{\partial^{n+m} f_i}{\partial x^n \, \partial y_{j_1} \dots \partial y_{j_m}}.$$

La forme de ces relations est la même que dans le cas d'une seule équation. En évaluant l'erreur dans les méthodes des différences finies, on remplace

$$f(x_n, y_n) - f(x_n, y(x_n)) = f_y(x_n, \tilde{y}_n) R_n$$

par

$$\mathbf{f}(x_n, \mathbf{y}_n) - \mathbf{f}(x_n, \mathbf{y}(x_n)) = f_y^n \mathbf{R}_n,$$

où f_y^n est une matrice d'éléments $\left.\frac{\partial f_i}{\partial y_j}\right|_{(x_n,\, y_n^i)}$; y_n^i est indicé i parce que la formule de Lagrange s'applique de façon indépendante à chaque ligne de la dernière égalité.

Toute erreur introduite au cours de l'intégration numérique peut admettre pour modèle la perturbation de la condition initiale imposée à l'intégration ultérieure (on l'a vu clairement au § 4 sur l'exemple de l'erreur dans la méthode de Runge-Kutta). Dans le cas d'une équation différentielle, la variation de la solution (la différence de deux solutions indéfiniment proches) vérifie l'équation aux variations $\eta' = f_y \eta$. Puisqu'il n'existe pas de raisons particulières pour que les solutions des approximations numériques soient plus proches que celles des équations différentielles, l'erreur se multiplie inévitablement par $\exp\left(\int f_y \, dx\right)$ comme dans la méthode de Runge-Kutta. Toutes nos constructions à l'aide des équations modèles $y' = My$ avaient pour but de mettre à l'index des schémas dans lesquels l'erreur accuse une croissance plus grande. S'il s'agit des systèmes, le système d'équations aux variations s'écrit $\boldsymbol{\eta}' = f_y \boldsymbol{\eta}$, f_y étant une matrice, et la contribution de l'erreur par pas à l'erreur totale est fonction du comportement de ses solutions. Si dans le cas d'une équation la partie principale de l'erreur résultant des méthodes des différences finies est proportionnelle à la solution de

$$z' = f_y z + y^{(m+1)}(x),$$

dans le cas d'un système, elle l'est à la solution de

$$\mathbf{z}' = f_y \mathbf{z} + \mathbf{y}^{(m+1)}(x).$$

Formellement écrits pour un système d'équations, les formules de Runge-Kutta et les schémas aux différences finies donnent lieu à une erreur du même ordre de grandeur que dans le cas d'une seule équation.

Considérons le problème modèle d'intégration du système linéaire à coefficients constants

$$\mathbf{y}' = A\mathbf{y}. \tag{1}$$

Supposons, afin de simplifier l'exposé, que la forme de Jordan de la matrice A est simple:

$$C^{-1}AC = \Lambda,$$

avec Λ une matrice diagonale. Les formules de calcul de la méthode de Runge-Kutta s'écrivent alors

$$\mathbf{k}_1 = hA\mathbf{y},$$
$$\cdots\cdots$$
$$\mathbf{k}_q = hA\left(\mathbf{y} + \sum_{i=1}^{q-1} \beta_{qi}\mathbf{k}_i\right),$$
$$\cdots\cdots$$
$$\mathbf{y}(x+h) \approx \mathbf{z}(h) = \mathbf{y} + \sum_{i=1}^{q} p_i\mathbf{k}_i.$$

En exprimant de proche en proche $\mathbf{k}_2, \ldots, \mathbf{k}_q$, $\mathbf{z}(h)$ en fonction de $\mathbf{y}$, on aboutit à

$$\mathbf{k}_1 = Ah\mathbf{y},$$
$$\mathbf{k}_2 = P_2(Ah)\,\mathbf{y},$$
$$\cdots\cdots\cdots\cdots$$
$$\mathbf{k}_q = P_q(Ah)\,\mathbf{y},$$
$$\mathbf{z}(h) = Q_q(Ah)\,\mathbf{y},$$

où P_l, Q_l sont des polynômes de degré l en la matrice Ah dont les coefficients dépendent des paramètres α_i, β_{ij}, p_i du procédé considéré. La relation

$$\mathbf{y}_j = Q_q(Ah)\,\mathbf{y}_{j-1} \tag{2}$$

est vraie à chaque pas d'intégration. On obtient par conséquent, après n pas, la valeur $\mathbf{y}_n = (Q_q(Ah))^n\mathbf{y}_0$. Cette égalité s'écrit

$$y_n = (Q_q(\lambda h))^n\, y_0$$

quand on intègre une seule équation $y' = \lambda y$.

Soient $\mathbf{e}_1, \ldots, \mathbf{e}_m$ les vecteurs propres de la matrice A et $\lambda_1, \ldots, \lambda_m$ ses valeurs propres associées. Etant donné que par hypothèse A est une matrice de structure simple, ces vecteurs forment une base; tout vecteur $\mathbf{y}_j$ est représenté de façon univoque dans cette base:

$$\mathbf{y}_j = \sum_{p=1}^{m} c_j^p\mathbf{e}_p.$$

Portons les expressions de $\mathbf{y}_n$, $\mathbf{y}_{n-1}$ en fonction des vecteurs de la base dans le second membre de (2). La multiplication du polynô-

me $S(A)$ en la matrice A par le vecteur propre $\mathbf{e}_p$ de A fournit

$$S(A)\,\mathbf{e}_p = S(\lambda_p)\,\mathbf{e}_p.$$

Donc,

$$Q_q(Ah)\,\mathbf{y}_{n-1} = \sum_{p=1}^{m} Q_q(Ah)\,c_{n-1}^p\,\mathbf{e}_p = \sum_{p=1}^{m} Q_q(\lambda_p h)\,c_{n-1}^p\,\mathbf{e}_p,$$

et

$$c_n^p = Q_q(\lambda_p h)\,c_{n-1}^p,$$

$$c_n^p = (Q_q(\lambda_p h))^n\,c_0^p, \tag{3}$$

$$\mathbf{y}_n = \sum_{p=1}^{m} (Q_q(\lambda_p h))^n\,c_n^p \mathbf{e}_p. \tag{4}$$

L'examen des relations (3), (4) nous autorise à faire les conclusions suivantes sur l'intégration du système linéaire (1) par la technique de Runge-Kutta : chaque composante de la solution proportionnelle à l'un des vecteurs propres de la matrice A est intégrée indépendamment des autres ; les formules de transformation des coefficients c_n^p coïncident alors avec les formules d'intégration numérique de l'équation $y' = \lambda_p y$ par les méthodes de Runge-Kutta.

Passons aux méthodes des différences finies. Pour le système (1) le système correspondant d'équations aux différences finies a la forme

$$\sum_{i=0}^{k} a_{-i}\mathbf{y}_{n-i} - h\sum_{i=0}^{k} b_{-i}A\mathbf{y}_{n-i} = 0\,; \tag{5}$$

posons comme toujours

$$\mathbf{y}_n = \sum_{p=1}^{m} c_n^p \mathbf{e}_p$$

et portons cette valeur dans (5) ; il vient

$$\sum_{p=1}^{m} \Big(\sum_{i=0}^{k} a_{-i}c_{n-i}^p - h\sum_{i=0}^{k} b_{-i}\lambda_p c_{n-i}^p\Big)\,\mathbf{e}_p = 0.$$

Par suite de l'indépendance linéaire des vecteurs $\mathbf{e}_p$, chaque parenthèse s'annule. On a donc

$$\sum_{i=0}^{k} a_{-i}c_{n-i}^p - h\sum_{i=0}^{k} b_{-i}\lambda_p c_{n-i}^p = 0.$$

Cette équation relativement à c_n^p coïncide avec l'équation aux différences obtenue en intégrant l'équation $y' = \lambda_p y$ par la méthode considérée.

Dans les deux cas (qu'il s'agisse des techniques de Runge-Kutta ou des méthodes des différences finies), les composantes de la solution proportionnelles aux vecteurs propres $\mathbf{e}_p$ sont intégrées indépendamment les unes des autres et les formules de calcul coïncident avec les formules d'intégration des équations scalaires $y' = \lambda_p y$. Si un pas d'intégration donné garantit une précision acceptable dans l'intégration de chaque équation $y'_p = \lambda_p y_p$, il en est de même pour le système dans son ensemble.

Les solutions const exp $(\lambda_p x)$ de $y' = \lambda_p y$ varient sensiblement sur un segment d'ordre de longueur $|\lambda_p|^{-1}$. L'intégration de cette composante de la solution avec une précision acceptable exige donc que $h \ll |\lambda_p|^{-1}$, ou encore $|\lambda_p| h \ll 1$. Pour intégrer toutes les composantes avec une précision élevée il faut que

$$\max_p |\lambda_p| h \ll 1. \tag{6}$$

Prenons maintenant le système non linéaire

$$\mathbf{y}' = \mathbf{f}(x, \mathbf{y}).$$

Au voisinage de la solution $\mathbf{Y}(x)$ cherchée son second membre est proche de celui du système linéaire

$$\mathbf{y}' = \mathbf{f}(x, \mathbf{Y}) + f_y(x, \mathbf{Y})(\mathbf{y} - \mathbf{Y}).$$

Sa solution est intégrée avec une précision forte s'il en est de même de la solution du système linéaire

$$\mathbf{z}' = f_y(x, \mathbf{Y})\,\mathbf{z}.$$

Au voisinage de tout point X ce système est approché par le système linéaire à coefficients constants

$$\mathbf{u}' = f_y(X, \mathbf{Y}(X))\,\mathbf{u}.$$

Ainsi, une condition nécessaire pour que la solution du système initial soit intégrée avec une précision assez élevée est $\max_p |\lambda_p(X)| h \ll$ $\ll 1$, où $\lambda_p(X)$ sont les valeurs propres de la matrice $f_y(X, \mathbf{Y}(X))$. Puisque cette condition résulte de différentes simplifications du système de départ, elle ne suffit pas pour assurer la proximité des solutions des problèmes discrétisé et différentiel.

Cependant, dans l'intégration de certaines classes de problèmes la condition (6) s'avère trop encombrante. De ces classes on peut donner la description qualitative suivante. Le système considéré possède une solution $\mathbf{y}_0(x)$ asymptotiquement stable. Parmi les solutions dans un voisinage de $\mathbf{y}_0(x)$, certaines tendent vers celle-ci relativement vite et d'autres relativement lentement. On recherche numériquement une solution se rapprochant lentement de $\mathbf{y}_0(x)$ et son allure ne présente de l'intérêt que pour des x suffisamment grands. On prend comme modèle le système de deux

équations linéaires $y' = Ay$ dotées des propriétés suivantes. Les deux valeurs propres λ_k de la matrice A sont négatives, de plus $|\lambda_1| \ll |\lambda_2|$. Il y a intérêt à résoudre ce système pour x tels que $|\lambda_2| x \gg 1$ et $|\lambda_1| x$ pas très important. La solution du problème examiné s'écrit

$$C_1 \exp(\lambda_1 x)\, e_1 + C_2 \exp(\lambda_2 x)\, e_2.$$

Les conditions imposées ci-dessus à x signifient que la deuxième composante $C_2 \exp(\lambda_2 x) e_2$ est négligeable contrairement à la première. Celle-ci sera intégrée avec une précision admissible si $|\lambda_1| h \ll 1$. Comme $|\lambda_1| x$ n'est pas très grand, on serait tenté de croire que le problème considéré se résoudra sans de grandes difficultés.

Cependant, le fait que l'intégration de la deuxième composante de la solution n'est assujettie à aucune condition de précision ne veut nullement dire qu'on peut lui appliquer n'importe quelle méthode.

Voyons par exemple les méthodes d'Adams explicites. Si la condition (10.15) relativement à $M = \lambda_2$ n'est pas réalisée, le schéma aux différences possède une racine parasite de module supérieur à 1 et la deuxième composante de la solution croît fortement. Il est donc intéressant de construire une méthode dans laquelle la deuxième composante se prête peut-être mal à l'intégration, mais s'amortit rapidement. Recourons à des méthodes du type de Runge-Kutta telles que, dans le cas de l'équation linéaire $y' = \lambda y$, l'équation caractéristique correspondante ait une racine caractéristique unique. Notons μ la quantité $Q_q(\lambda h)$.

Pour la méthode d'Euler

$$z(h) = y + k_1 = (1 + \lambda h)\, y.$$

Ainsi, $\mu = 1 + \lambda h$ et la condition $|\mu| \leqslant 1$, lorsque $\lambda < 0$, n'est plus observée dès que $|\lambda| h > 2$.

Pour la méthode (2.7) exacte à l'ordre 2

$$k_1 = \lambda h y,$$

$$k_2 = hf\left(x + \frac{h}{2},\ y + \frac{k_1}{2}\right) = \lambda h\left(y + \frac{k_1}{2}\right) = \lambda h\left(y + \frac{\lambda h y}{2}\right),$$

$$z(h) = y_0 + k_2 = \left(1 + \lambda h + \frac{\lambda^2 h^2}{2}\right) y,$$

$$\mu = 1 + \lambda h + \frac{\lambda^2 h^2}{2},$$

et la condition $|\mu| \leqslant 1$ pour $\lambda < 0$ n'est plus réalisée quand $|\lambda| h > 2$. Pour la méthode (2.6) exacte à l'ordre 2

$$k_1 = \lambda h y,$$

$$k_2 = \lambda h (y + k_1) = \lambda h (y + \lambda h y),$$

$$z(h) = y + \frac{1}{2}(k_1 + k_2) = \left(1 + \lambda h + \frac{\lambda^2 h^2}{2}\right) y,$$

avec la même condition $|\mu| \leqslant 1$ pour $\lambda < 0$.

Dans le cas des schémas exacts au quatrième ordre (2.14) et (2.15)

$$\mu = 1 + \lambda h + \frac{\lambda^2 h^2}{2} + \frac{\lambda^3 h^3}{6} + \frac{\lambda^4 h^4}{24};$$

$|\mu| \leqslant 1$ lorsque $\lambda < 0$ si $|\lambda h| \leqslant \varkappa \approx 2{,}8$. Les méthodes de Runge-Kutta ne jouent donc pour le problème en question que dans la condition du type

$$|\lambda_2| h \leqslant \text{const}.$$

PROBLÈME. Démontrer que pour les méthodes de Runge-Kutta avec $q = s$ on a toujours

$$z(h) = \left(\sum_{k=0}^{s} \frac{\lambda^k h^k}{k!}\right) y$$

(pour la définition de q et s voir § 2).

La mise en œuvre de la méthode d'Adams implicite exacte à l'ordre 2 nous fournit une solution plus acceptable de ce problème:

$$y_n - y_{n-1} = \frac{h}{2}(f_n + f_{n-1}).$$

Substituant λy à f nous trouvons $y_n = \mu y_{n-1}$, avec $\mu = \dfrac{1 + \frac{\lambda h}{2}}{1 - \frac{\lambda h}{2}}$. Lorsque $\lambda = \alpha + i\beta$.

$$\mu = \frac{1 + \frac{\alpha h}{2} + i\frac{\beta h}{2}}{1 - \frac{\alpha h}{2} - i\frac{\beta h}{2}}, \qquad |\mu| = \frac{\sqrt{\left(1 + \frac{\alpha h}{2}\right)^2 + \left(\frac{\beta h}{2}\right)^2}}{\sqrt{\left(1 - \frac{\alpha h}{2}\right)^2 + \left(\frac{\beta h}{2}\right)^2}}.$$

Si $\alpha = \operatorname{Re}\lambda < 0$, on a

$$\left(1 + \frac{\alpha h}{2}\right)^2 + \left(\frac{\beta h}{2}\right)^2 < \left(1 - \frac{\alpha h}{2}\right)^2 + \left(\frac{\beta h}{2}\right)^2,$$

et donc $|\mu| < 1$. Le schéma assure, on le voit, la non-croissance des composantes de la solution correspondant à tous les λ avec $\operatorname{Re}\lambda < 0$. L'intégration de la première composante s'effectue avec une exactitude acceptable si $|\lambda_1| h \ll 1$. La non-croissance de la deuxième composante ne signifie cependant pas sa petitesse.

On a l'égalité

$$\mu_2 = \frac{1 + \frac{\lambda_2 h}{2}}{1 - \frac{\lambda_2 h}{2}} = -\frac{1 + \frac{2}{\lambda_2 h}}{1 - \frac{2}{\lambda_2 h}};$$

il en résulte que pour $\lambda_2 < 0$ la quantité $|\mu_2|$ se rapproche de 1 par valeurs inférieures avec l'augmentation de $|\lambda_2| h$ et la deuxième composante s'amortit, par conséquent, lentement. Son amortissement rapide impose une certaine contrainte à la borne supérieure du pas h, qui est néanmoins plus lâche que $|\lambda_2| h \ll 1$. Pour un système linéaire à coefficients variables l'analogue du

dernier schéma sont les schémas d'ordre d'exactitude 2:

$$\frac{y_n - y_{n-1}}{h} - \frac{A(x_n)\, y_n + A(x_{n-1})\, y_{n-1}}{2} - F_{n-\frac{1}{2}} = 0,$$
$$\frac{y_n - y_{n-1}}{h} - A\left(x_{n-\frac{1}{2}}\right) \frac{y_n + y_{n-1}}{2} - F_{n-\frac{1}{2}} = 0, \tag{7}$$

avec

$$F_{n-\frac{1}{2}} = \frac{f(x_n) + f(x_{n-1})}{2}.$$

On préférera parfois à cette dernière valeur

$$F_{n-\frac{1}{2}} = f\left(x_{n-\frac{1}{2}}\right) \quad \text{ou} \quad F_{n-\frac{1}{2}} = \frac{1}{h} \int_{x_{n-1}}^{x_n} f(x)\, dx.$$

Dans les formules des deux types il faut à chaque pas inverser une matrice dont la dimension est celle du système. Si l'ordre du système est élevé, on perfectionne la méthode pour ne pas avoir à inverser une matrice à chaque pas.

Voici un autre procédé de résolution des problèmes du genre. Prenons, pour commencer, le système $y' = Ay$, $A = \text{const}$. L'intégration numérique s'effectue à l'aide de la formule

$$y(x_n + h) = \varphi(h)\, y(x_n),$$

où

$$\varphi(h) = \exp(Ah).$$

La fonction matricielle $\varphi(H)$ vérifie

$$\varphi(H) = \varphi\left(\frac{H}{2}\right) \varphi\left(\frac{H}{2}\right), \tag{8}$$

et on la calcule donc comme suit. On se donne un certain l, on calcule la matrice $\varphi\left(\frac{h}{2^l}\right)$, puis, moyennant l'égalité (8), on trouve successivement les matrices

$$\varphi\left(\frac{h}{2^{l-1}}\right), \ldots, \varphi\left(\frac{h}{2}\right), \varphi(h).$$

$\varphi\left(\frac{h}{2^l}\right)$ peut être remplacée par une somme partielle de sa série de Taylor

$$\varphi\left(\frac{h}{2^l}\right) \approx \sum_{i=0}^{s} \frac{1}{i!} \left(\frac{Ah}{2^l}\right)^i. \tag{9}$$

On remarquera que pour $q = s \leqslant 4$ le calcul par la formule (9) équivaut à intégrer le système d'équations matricielles $Y' = AY$ par la méthode de Runge-Kutta d'ordre s avec le pas $h/2^l$ et la condition initiale $Y(0) = E$, où E est la matrice unité.

La petitesse de l'erreur dans (9) pour s faibles dépend essentiellement de la condition $\| A \| h/2^l \ll 1$. Aussi la quantité 2^l peut s'avérer assez élevée.

Pour éviter une accumulation indésirable de l'erreur de calcul, introduisons (par analogie avec le § 9, chap. IV) la fonction matricielle $\psi(h) = \varphi(h) - E$. On calcule

$$\psi\left(\frac{h}{2^l}\right) \approx \sum_{i=1}^{s} \frac{1}{i!}\left(\frac{Ah}{2^l}\right)^i,$$

ensuite

$$\psi\left(\frac{h}{2^{l-1}}\right), \ldots, \psi\left(\frac{h}{2}\right), \psi(h),$$

par la formule de récurrence

$$\psi(H) = \psi\left(\frac{H}{2}\right)\left(2E + \psi\left(\frac{H}{2}\right)\right);$$

puis on trouve

$$\varphi(H) = E + \psi(H).$$

Considérons maintenant le système non homogène

$$\mathbf{y}' = A\mathbf{y} + \mathbf{b}, \qquad A,\ \mathbf{b} = \text{const}.$$

La différence $\mathbf{u}(x) = \mathbf{y}(x) - \mathbf{y}(x_n)$ satisfait au système $\mathbf{u}' = A\mathbf{u} + \mathbf{c}$, avec $\mathbf{c} = A\mathbf{y}(x_n) + \mathbf{b}$. Faisons intervenir la matrice

$$\omega(H) = \int_0^H \exp(A(H-t))\,dt.$$

On vérifie aisément que $\mathbf{u}(x_n + h) = \omega(h)\mathbf{c}$, et la matrice $\omega(h)$ est telle que

$$\omega(H) = \omega\left(\frac{H}{2}\right)\left(2E + A\omega\left(\frac{H}{2}\right)\right); \tag{10}$$

en outre

$$\omega(H) = \sum_{i=1}^{\infty} \frac{1}{i!} A^{i-1} H^i.$$

C'est pourquoi on calcule

$$\omega\left(\frac{h}{2^l}\right) \approx \sum_{i=1}^{s} \frac{1}{i!} A^{i-1}\left(\frac{h}{2^l}\right)^i,$$

ensuite

$$\omega\left(\frac{h}{2^{l-1}}\right), \ldots, \omega\left(\frac{h}{2}\right), \omega(h)$$

à l'aide de la relation (10) et on pose

$$\mathbf{y}(x_n + h) \approx \mathbf{y}(x_n) + \omega(h)(A\mathbf{y}(x_n) + \mathbf{b}).$$

S'il s'agit de systèmes de forme générale, linéaires ou non, on approche le système considéré sur chaque segment $[x_n, x_{n+1}]$ par le système

$$\mathbf{y}' = A\mathbf{y} + \mathbf{b}, \quad A, \mathbf{b} = \text{const.}$$

§ 12. Quelques questions générales

Pour faire démarrer les calculs suivant un schéma aux différences finies d'ordre $k > 1$ on doit connaître les valeurs approchées $y_0, \ldots, y_{k-1}$ des $y(x_0), \ldots, y(x_{k-1})$. La valeur y_0 est donnée par la condition initiale. Tout ce que nous savons des méthodes numériques nous oblige à rechercher $y_1, \ldots, y_{k-1}$ par une méthode de Runge-Kutta à pas séparés. Or, d'habitude on s'adresse à des procédés un peu différents.

Prenons un nombre q fixe et un nombre j quelconque tels que $0 < j \leqslant q$. Remplaçons à droite dans

$$y(x_j) - y(x_0) = \int_{x_0}^{x_j} y'(x)\,dx$$

$y'(x)$ par le polynôme d'interpolation de Lagrange basé sur $x_0, \ldots, x_q$:

$$y'(x) = \sum_{i=0}^{q} y'(x_i) \prod_{p \neq i} \frac{x - x_p}{x_i - x_p} + r(x),$$

$$r(x) = y'(x_0; \ldots; x_q; x) \prod_{p=0}^{q} (x - x_p);$$

il vient

$$y(x_j) - y(x_0) = h \sum_{i=0}^{q} c_{ji} y'(x_i) + R_j, \tag{1}$$

où

$$c_{ji} = \frac{1}{h} \int_{x_0}^{x_j} \prod_{p \neq i} \frac{x - x_p}{x_i - x_p}\,dx = \int_0^j \prod_{p \neq i} \frac{t - p}{i - p}\,dt,$$

$$R_j = \int_{x_0}^{x_j} r(x)\,dx.$$

La quantité $|y'(x_0; \ldots; x_q; x)|$ est majorée par

$$\max_{[x_0,\, x_q]} \left| \frac{y^{(q+2)}(x)}{(q+1)!} \right|.$$

En changeant de variable $x=th$ dans l'expression de R_j et en évaluant en module fonction à intégrer, on obtient

$$|R_j| \leqslant c_j \max_{[x_0,\, x_q]} |y^{q+2}(x)|\, h^{q+2}, \tag{2}$$

$$c_j = \int_0^j \frac{\left| \prod_{p=0}^{q} (t-p) \right|}{(q+1)!}\, dt.$$

Acceptons pour approximations des $y(x_j)$ la solution du système

$$y_j - y_0 = h \sum_{i=0}^{q} c_{ji} f(x_i, y_i). \tag{3}$$

Retranchons (1) de (3); ceci étant, utilisons le fait que la différence

$$f(x_i, y_i) - y'(x_i) = f(x_i, y_i) - f(x_i, y(x_i))$$

peut, en vertu de la formule de Lagrange, se mettre sous la forme $l_i r_i$ avec

$$r_i = y_i - y(x_i), \quad l_i = f_y(x_i, \widetilde{y}_i),$$

$\widetilde{y}_i$ étant entre y_i et $y(x_i)$. Nous trouvons

$$r_j - r_0 = h \sum_{i=0}^{q} c_{ji} l_i r_i - R_j. \tag{4}$$

Posons

$$L = \max_{x_0 \leqslant x \leqslant x_q} |f_y|,$$

$$c = \max_j \sum_{i=0}^{q} |c_{ji}|.$$

L'égalité (4) entraîne

$$|r_j| \leqslant |r_0| + cLh \max_{0 \leqslant i \leqslant q} |r_i| + |R_j|, \tag{5}$$

d'où

$$\max_{0 \leqslant j \leqslant q} |r_j| \leqslant |r_0| + \max_{0 \leqslant j \leqslant q} |R_j| + cLh \max_{0 \leqslant i \leqslant q} |r_i|.$$

Le report du dernier terme dans le deuxième membre donne

$$(1 - cLh) \max_{0 \leqslant i \leqslant q} |r_i| \leqslant |r_0| + \max_{0 \leqslant i \leqslant q} |R_j|.$$

Soit ensuite h tellement petit que $cLh < 1$. Alors

$$\max_{0 \leqslant i \leqslant q} |r_i| \leqslant \frac{|r_0| + \max\limits_{0 \leqslant i \leqslant q} |R_j|}{1 - cLh}.$$

Si la valeur donnée y_0 est supposée exacte, i.e. $r_0 = 0$, il en résulte l'estimation de l'erreur

$$\max_{0 \leqslant i \leqslant q} |r_i| = O(h^{q+2}) \tag{6}$$

qui se prête à une amélioration dans certains cas. Si q est pair par exemple, les points d'interpolation sont symétriques par rapport à $x_0 + \frac{q}{2}h$, point médian du segment $[x_0, x_q]$. L'erreur R_q dans la formule de quadrature

$$y(x_q) - y(x_0) \approx h \sum_{k=0}^{q} c_{qk} y'(x_k)$$

est alors (nous l'avons montré dans la partie de ce livre consacrée à l'intégration numérique) une quantité de l'ordre de $O(h^{q+3})$. En portant cette estimation dans le deuxième membre de l'inégalité (5) pour $j = q$, et en nous servant de (6) nous obtenons $r_q = O(h^{q+3})$. On utilise cette circonstance dans l'intégration à l'aide des formules d'Adams où il importe tout particulièrement que y_q, q maximal, soit précis au possible. En effet, dans l'intégration suivant la formule d'Adams d'ordre k la contribution de l'erreur sur les conditions initiales est

$$O\left(|y_{k-1} - y(x_{k-1})| + h \sum_{j=0}^{k-2} y_j - y(x_j)|\right).$$

Dans le cas non linéaire, le système (3) est en général irrésoluble par rapport aux inconnues $y_1, \ldots, y_q$. On le résout par la méthode itérative simple ou par une méthode du type (7.1.8):

$$y_j^{m+1} - y_0 = h \sum_{i=0}^{j-1} c_{ji} f(x_i, y_i^{m+1}) + h \sum_{i=j}^{q} c_{ji} f(x_i, y_i^m).$$

Les dérivées par rapport aux variables y_j à droite sont multipliées par h, ce qui fait que, lorsque h est faible, les processus itératifs considérés réalisent une contraction et, partant, convergent. Le choix de l'approximation de départ, du nombre q et du nombre d'itérations est souvent dicté non pas par des considérations purement mathématiques, mais par des circonstances liées entre autres au volume de la mémoire rapide, à l'ordre type des systèmes proposés.

Etudions plus en détail le calcul par pas des valeurs y_n dans la méthode d'Adams implicite. D'habitude la méthode d'Adams implicite

$$y_n - y_{n-1} - h \sum_{i=0}^{k} b_{-i} f(x_{n-i}, y_{n-i}) = 0 \tag{7}$$

est employée conjointement avec la méthode explicite

$$y_n - y_{n-1} - h \sum_{i=1}^{k+1} \overline{b}_{-i} f(x_{n-i}, y_{n-i}) = 0.$$

L'ordre d'approximation des deux formules est par construction en $O(h^{k+1})$. Avec les méthodes implicites on a à résoudre une équation non linéaire ou un système de telles équations, ce qui semble entraîner des complications inutiles par comparaison avec les méthodes explicites. D'autre part, la partie principale de leur erreur par pas comprend un facteur constant un peu moins grand, et, dans le cas de l'équation linéaire, la contrainte par pas du type (10.15) qui garantit pour $f_y < 0$ l'absence de racines de module supérieur à 1 de l'équation (7.3) s'avère moins sévère. C'est sans doute cette dernière circonstance qui a déterminé la vogue des méthodes implicites.

Connaissant y_{n-1} on détermine l'approximation d'ordre zéro de la valeur y_n :

$$y_n^0 = y_{n-1} + h \sum_{i=1}^{k+1} \overline{b}_{-i} f(x_{n-i}, y_{n-i});$$

ensuite on effectue plusieurs itérations à l'aide de

$$y_n^{m+1} = y_{n-1} + h \sum_{i=1}^{k} b_{-i} f(x_{n-i}, y_{n-i}) + h b_0 f(x_0, y_n^m). \tag{8}$$

La dérivée par rapport à y_n à droite vaut $nb_0 f_y$, c'est pourquoi à chaque pas du processus, l'erreur sur les valeurs y_n^m acquiert, par rapport à celle sur la solution exacte de (7), un facteur en h. Vu la longueur de l'analyse détaillée de cette erreur nous nous en abstiendrons. Il se trouve que pour que la partie principale de l'erreur soit la même que dans le cas de la solution exacte de (7) par rapport à l'inconnue y_n, une seule itération d'après la formule (8) suffit. Puisque le calcul du deuxième membre est d'ordinaire notablement plus laborieux que les autres opérations auxiliaires, on évalue naturellement le travail imposé par la méthode par le nombre de calculs de ce membre. Si à chaque pas on effectue une itération et qu'ensuite on détermine la valeur $f(x_n, y_n^1)$ nécessaire aux calculs suivants, il faudra deux fois plus de travail que dans la méthode d'Adams explicite. On l'évite en posant chaque fois

$$f(x_n, y_n^1) \approx f(x_n, y_n^0).$$

Ce faisant, le mode opératoire s'avère d'un même coût que la méthode d'Adams explicite et la partie principale de son erreur est égale à celle de la méthode implicite. Dans la pratique, cette variante de la méthode implicite risque de faire croître (par comparaison au procédé implicite même) le pas jusqu'à une valeur telle qu'il y ait influence des racines parasites. Aussi impose-t-on parfois un fractionnement supplémentaire du pas par rapport à la recherche de la valeur exacte y_n à partir de l'équation sous forme implicite (7). Ceux qui rédigent les programmes usuels de la méthode d'Adams s'arrangent pour pouvoir à la fois effectuer une fragmentation supplémentaire du pas et augmenter à volonté le nombre d'itérations par pas compte tenu des propriétés statistiques des problèmes à résoudre.

On a parfois à intégrer numériquement des équations différentielles telles que la dérivée de la solution ne s'exprime pas en fonction de cette solution et de l'argument ; pour fixer les idées prenon le cas scalaire

$$F(x, y, y') = 0.$$

En résolvant cette équation par rapport à y', on aurait obtenu

$$y' = f(x, y).$$

Une première voie possible est d'appliquer formellement les méthodes de Runge-Kutta ou des différences finies. On calcule f pour tout x et tout y en résolvant numériquement

$$F(x, y, f) = 0.$$

Une bonne approximation initiale de la valeur f cherchée s'obtient par interpolation des valeurs déjà trouvées ; aussi le nombre néces-

saire d'itérations (c'est le cas de la méthode de la sécante, par exemple) est d'ordinaire peu élevé.

Pour les méthodes implicites, on gagne parfois à résoudre d'emblée le système

$$F(x_n, y_n, f_n) = 0,$$

$$\sum_{i=0}^{k} a_{-i} y_{n-i} - h \sum_{i=0}^{k} b_{-i} f_{n-i} = 0$$

relativement aux inconnues y_n, f_n.

Le développement de la solution en série de Taylor s'avère parfois (mais assez rarement) un procédé utile; la dérivation par rapport à x de l'équation de départ donne

$$\frac{d}{dx}(F(x, y, y')) = F_x(x, y, y') + F_y(x, y, y')\, y' + F_{y'}(x, y, y')\, y'' = 0,$$

$$\frac{d^2}{dx^2}(F(x, y, y')) = \ldots = 0,$$

et ainsi de suite. Si l'on connaît $y'(x_0)$, ces relations permettent de déterminer explicitement les valeurs $y''(x_0)$, ...

La première des relations ci-dessus se récrit

$$y'' = g(x, y, y').$$

D'où une troisième voie conduisant à la solution. À l'aide de

$$F(x_0, y_0, y'(x_0)) = 0$$

on obtient $y'(x_0)$, puis on intègre numériquement l'équation

$$y'' = g(x, y, y').$$

Quelques mots sur le changement du pas d'intégration. Nous avons étudié les méthodes des différences finies en supposant le pas $h = x_n - x_{n-1}$ constant. Or, il faut souvent intégrer avec un pas variable. Ainsi, les programmes standard d'intégration par la méthode d'Adams avec une recherche automatique du pas réalisent de temps en temps une duplication ou une dichotomie du pas selon le degré de régularité de la solution. On peut doubler le pas en négligeant un point d'intégration sur deux et en acceptant les valeurs restantes des inconnues comme valeurs initiales. Cependant, l'augmentation et le fractionnement du pas sont généralement effectués par les formules d'intégration sur un réseau irrégulier. On construit ces formules par la méthode des coefficients indéterminés ou par le procédé suivant. Supposons connues les approximations y_{n-1}, ..., y_{n-k} des valeurs $y(x_{n-1})$, ..., $y(x_{n-k})$ de la solution. Substituons à $y'(x)$ dans

$$y(x_n) - y(x_{n-1}) = \int_{x_{n-1}}^{x_n} y'(x)\, dx$$

la valeur du polynôme d'interpolation basé sur les points $x_{n-1}, \ldots, x_{n-k}$. Nous trouvons

$$y(x_n) - y(x_{n-1}) \approx \sum_{i=1}^{k} b_{-i}(x_n, \ldots, x_{n-k})\, y'(x_{n-i}).$$

Il lui correspond la formule explicite

$$y_n - y_{n-1} = \sum_{i=1}^{k} b_{-i}(x_n, \ldots, x_{n-k})\, f(x_{n-i}, y_{n-i}).$$

Si l'on interpole aux points $x_n, \ldots, x_{n-k}$, on obtient la formule implicite

$$y_n - y_{n-1} = \sum_{i=0}^{k} b_{-i}(x_n, \ldots, x_{n-k})\, f(x_{n-i}, y_{n-i}).$$

Discutons un autre problème de caractère général. En évaluant l'erreur on a toujours supposé qu'était réalisée une condition très sévère, à savoir $|f_y| \leqslant L$ pour $x_0 \leqslant x \leqslant x_0 + X$ et quels que soient y.

En réalité, toutes les estimations ci-dessus de l'erreur sont vraies pour h et δh^{-1} suffisamment faibles si $|f_y| \leqslant L$ dans le β-voisinage de la solution:

$$x_0 \leqslant x \leqslant x_0 + X, \quad |y - y(x)| \leqslant \beta, \text{ où } \beta > 0.$$

Sous l'hypothèse $|f_y| \leqslant L$ avec y arbitraire, on a une estimation d'erreur, disons (8.13), dont le deuxième membre tend vers zéro avec h et δh^{-1}. Supposons ces dernières quantités si petites que le deuxième membre de l'estimation est au plus β en tous les points du segment d'intégration. Tous les points y_n et, partant, tous les points intermédiaires $\bar{y}_n$ appartiennent alors, eux aussi, au β-voisinage de la solution, et en évaluant l'erreur on ne se sert pas de l'information sur les valeurs de la fonction extérieures à ce voisinage. Donc, pour de tels h et δh^{-1} on a l'estimation (8.13) sous la condition $|f_y| \leqslant L$ dans le β-voisinage de la solution. On peut également lever la contrainte $|f_y| \leqslant L$, pour tous les y, dans les autres cas, par exemple en étudiant la convergence d'un processus itératif par les formules implicites. On se débarrasse de même de la condition assujettissant toutes les quantités figurant dans l'expression du reste à être bornées pour n'importe quel y.

Revenons une fois de plus à l'intégration des équations et des systèmes sur des intervalles temporels importants. Pour résoudre le système d'équations aux variations $\eta' = f_y \eta$ écrivons

$$\eta(x_2) = \exp\left(\int_{x_1}^{x_2} f_y(x, y(x))\, dx\right) \eta(x_1).$$

De nombreux problèmes asymptotiquement stables uniformément en x_1, x_2 dans les limites $x_0 \leqslant x_1 \leqslant x_2 \leqslant \infty$ satisfont à l'estimation

$$\left\| \exp\left(\int_{x_1}^{x_2} f_y(x, y(x))\, dx\right) \right\| \leqslant B \exp(-b(x_2 - x_1)), \quad b > 0. \tag{9}$$

Supposons que dans le domaine $\|y - y(x)\| \leqslant \beta$, $\beta > 0$, $x_0 \leqslant x < \infty$, les dérivées du deuxième membre du système dont les expressions figurent dans l'estimation du reste, en particulier f_y, sont uniformément bornées. C'est vrai pour de nombreux problèmes dont les solutions pour des x grands deviennent périodiques ou presque périodiques. Dans le cas des méthodes de Runge-Kutta ou d'Adams avec un pas d'intégration h constant suffisamment faible et δh^{-1}

petit, l'erreur est évaluée dans ces problèmes par

$$| y_n - y(x_n) | \leqslant B_1 (h^s + \delta h^{-1}) + B_2 R \exp(-b(x_n - x_0)), \qquad (10)$$

s étant l'ordre d'approximation, R l'estimation de l'erreur sur les conditions initiales, δ le maximum des normes des erreurs δ_n commises dans les équations discrétisées (4.1), $b > 0$, b, B_1, B_2 ne dépendant pas de x_n.

Nous avons déjà obtenu une telle estimation (voir 4.11) dans la méthode de Runge-Kutta appliquée à une seule équation.

Servons-nous de l'intégration par une méthode de Runge-Kutta d'ordre s de l'équation linéaire

$$y' = a_0(x)\, y + a_1(x)$$

pour montrer pourquoi, dans des évaluations pareilles, le comportement du deuxième membre loin de la solution est sans importance.

Dans ce cas $f_y = a_0(x)$ et la condition (9) prend la forme

$$\exp\left(\int_{x_1}^{x_2} a_0(x)\, dx\right) \leqslant B \exp(-b(x_2 - x_1)), \quad b > 0. \qquad (11)$$

On suppose en outre que

$$\max_{0 \leqslant i \leqslant s,\ p=0,\ 1} \sup_{x_0 \leqslant x < \infty} | a_p^{(i)}(x) | \leqslant A < \infty.$$

Dans la condition (11), on a, selon (4.6),

$$| R_n | \leqslant \sum_{j=1}^{n} B \exp(-b(n-j)h) | \omega_j | + B \exp(-bnh) | R_0 | \leqslant$$

$$\leqslant \frac{B}{1 - \exp(-bh)} \max_{0 \leqslant j \leqslant n} | \omega_j | + B | R_0 | \exp(-bnh). \qquad (12)$$

Il découle de

$$y(x) = y(x_0) \exp\left(\int_{x_0}^{x} a_0(x)\, dx\right) + \int_{x_0}^{x} b(t) \exp\left(\int_{t}^{x} a_0(\tau)\, d\tau\right) dt$$

et de l'inégalité (11) que $| y(x) | \leqslant Y$ est uniformément borné lorsque $x_0 \leqslant x < \infty$.

Fixons un $\beta > 0$; montrons que lorsque $h \leqslant h_0(\beta)$, $\dfrac{\delta}{h} \leqslant \gamma(\beta)$, $| R_0 | \leqslant R(\beta)$, $h(\beta)$, $\gamma(\beta)$, $R(\beta) > 0$, toutes les valeurs y_n sont dans le domaine $| y_n - y(x_n) | \leqslant \beta$ et on a l'estimation d'erreur désirée.

Evaluons d'abord l'erreur sous l'hypothèse $| y_n - y(x_n) | \leqslant \beta$ quel que soit n. Alors $| y_n | \leqslant Y + \beta$ pour tout n. La quantité y_n et les dérivées $a_p^{(i)}(x)$ étant uniformément bornées, on en conclut que toutes les erreurs par pas

$$\rho_{j+1} = y_{j+1} - y_j(x_{j+1})$$

vérifient

$$| \rho_{j+1} | \leqslant B'(\beta)\, h^{s+1}.$$

Alors

$$| \omega_j | \leqslant B'(\beta)\, h^{s+1} + \delta$$

avec j quelconque, et (12) entraîne l'estimation

$$|R_n| < \frac{B}{1-\exp(-bh)} (B'(\beta) h^{s+1} + \delta) + B|R_0| \exp(-bnh). \qquad (13)$$

Par suite de (13) on obtient l'estimation

$$|R_n| \leqslant \frac{Bh}{1-\exp(-bh)} \left(B'(\beta) h^s + \frac{\delta}{h}\right) + B|R_0| \qquad (14)$$

uniforme en n. Le facteur $\frac{Bh}{1-\exp(-bh)}$ décroît de façon monotone en tendant vers la limite $\frac{B}{b} > 0$ lorsque $h \to 0$. C'est pourquoi on peut indiquer $h(\beta)$, $\gamma(\beta)$, $R(\beta) > 0$ tels que le deuxième membre de la dernière inégalité est au plus β pour

$$h \leqslant h(\beta), \quad \frac{\delta}{h} \leqslant \gamma(\beta), \quad |R_0| \leqslant R(\beta). \qquad (15)$$

LEMME. *Dans la condition* (15) *tous les* n *vérifient l'estimation* (13).

On le démontre par récurrence sur n. Pour $n = 0$ l'affirmation du lemme est vraie en vertu du choix même des quantités $h(\beta)$, $\gamma(\beta)$, $R(\beta)$. Supposons l'estimation (13) juste pour tous les $n \leqslant n_0$. Alors $|y_j - y(x_j)| \leqslant \beta$ lorsque $j \leqslant n_0$; selon les constructions ci-dessus on en tire $|\rho_j| \leqslant B'(\beta) h^{s+1}$ avec $j \leqslant n_0 + 1$. Il en résulte, compte tenu de (12), que l'estimation (13) est également vraie pour $n = n_0 + 1$, ce qui démontre le lemme.

Dans l'exemple considéré y_n sont uniformément bornés, ainsi que $a_0(x_n)$ et $a_1(x_n)$; il en découle la même propriété des erreurs d'arrondi δ_n, et toutes les évaluations effectuées ont donc un sens.

§ 13. Formules d'intégration numérique des équations du deuxième ordre

En introduisant des fonctions inconnues, on ramène les équations différentielles d'ordre supérieur à 1 et leurs systèmes à des systèmes d'équations du premier ordre. Donc, sur le plan formel, la résolution numérique du problème de Cauchy pour les équations d'ordre supérieur pourrait être considérée comme achevée.

Or, les méthodes de résolution adaptées spécialement à des équations d'ordre supérieur s'avèrent souvent plus puissantes. En élaborant ces méthodes il convient de ne pas perdre de vue que les systèmes d'équations d'ordre supérieur de type spécial se rencontrent fréquemment et que si l'on tient compte de leurs particularités on augmenterait d'avantage l'efficacité des méthodes en question. Ainsi, la plupart des problèmes de mécanique céleste se réduisent à l'intégration des systèmes d'équations

$$\mathbf{y}'' = \mathbf{f}(\mathbf{y}).$$

Considérons une classe un peu plus vaste d'équations

$$\mathbf{y}'' = \mathbf{f}(x, \mathbf{y}).$$

Comme d'habitude on raisonne sur une seule équation, le résultat se généralisant automatiquement à un système. Les méthodes d'intégration les plus répandues sont en l'occurrence les *méthodes de Störmer*: *explicite*

$$y_{n+1}-2y_n+y_{n-1}=h^2\sum_{i=0}^{k} b_{-i}f(x_{n-i},\ y_{n-i}) \tag{1}$$

et *implicite*

$$y_{n+1}-2y_n+y_{n-1}=h^2\sum_{i=-1}^{k} b_{-i}f(x_{n-i},\ y_{n-i}). \tag{2}$$

Ces méthodes s'obtiennent également par la technique des coefficients indéterminés en exigeant que la différence

$$\frac{y(x_{n+1})-2y(x_n)+y(x_{n-1})}{h^2}-\sum_i b_{-i}y''(x_{n-i})$$

soit de plus grand ordre en h. On suppose d'ordinaire $x_{j+1}-x_j\equiv h$.

Indiquons un mode opératoire explicite pour obtenir de telles formules. Substituons dans

$$\frac{y(x_{n+1})-2y(x_n)+y(x_{n-1})}{h^2}$$

le développement taylorien en x_n de $y(x)$ avec un reste sous forme intégrale exprimé par la dérivée seconde:

$$y(x)=y(x_n)+y'(x_n)(x-x_n)+\int_{x_n}^{x}(x-t)\,y''(t)\,dt.$$

Nous obtenons l'égalité

$$\frac{y(x_{n+1})-2y(x_n)+y(x_{n-1})}{h^2}=\int_{x_{n-1}}^{x_{n+1}} G(x)\,y''(x)\,dx,$$

$$G(x)=\begin{cases}\dfrac{x_{n+1}-x}{h^2} & \text{pour } x_n\leqslant x\leqslant x_{n+1},\\[2ex] \dfrac{x-x_{n-1}}{h^2} & \text{pour } x_{n-1}\leqslant x\leqslant x_n.\end{cases}$$

Remplaçons $y''(x)$ par le polynôme d'interpolation de Newton basé sur $x_n, \ldots, x_{n-k}$:

$$y''(x)=y''(x_n)+\frac{\nabla y''(x_n)}{h}(x-x_n)+\ldots+\frac{\nabla^k y''(x_n)}{h^k k!}(x-x_n)\ \ldots$$

$$\ldots(x-x_{n-k+1})+y''(x_n\,;\ \ldots\,;\ x_{n-k}\,;\ x)(x-x_n)\ \ldots\ (x-x_{n-k}).$$

Après avoir intégré nous trouvons

$$\frac{y(x_{n+1})-2y(x_n)+y(x_{n-1})}{h^2}=\sum_{p=0}^{k}\beta_p\nabla^p y''(x_n)+r_n^k,$$

où

$$\beta_p=\frac{1}{p!}\left(\int_0^{+1}(1-t)\,t\,\ldots\,(t+(p-1))\,dt+\int_0^{-1}(-1-t)\,t\,\ldots\,(t+(p-1))\,dt\right), \quad (3)$$

$$r_n^k=\int_{x_{n-1}}^{x_{n+1}}G(x)\,y''(x_n;\,\ldots;\,x_{n-k};\,x)(x-x_n)\,\ldots\,(x-x_{n-k})\,dx.$$

Dans l'expression du reste sortons de sous le signe d'intégration la valeur $y''(x_n;\,\ldots;\,x_{n-k};\,x)$ en un point intermédiaire. Il vient

$$r_n^k=\overline{\beta}_{k+1}y''(x_n;\,\ldots;\,x_{n-k};\,\widetilde{x})\,h^{k+1};$$

la substitution de $\frac{1}{(k+1)!}\,y^{(k+3)}(\overline{x}_n)$, expression de la différence divisée par la dérivée selon (2.5.4), à $y''(x_n;\,\ldots;\,x_{n-k};\,x)$ donne

$$r_n^k=\frac{\overline{\beta}_{k+1}}{(k+1)!}\,y^{(k+3)}(\overline{x}_n)\,h^{k+1}. \quad (4)$$

On vérifie immédiatement que $\frac{\overline{\beta}_{k+1}}{(k+1)!}=\beta_{k+1}$, β_{k+1} étant déterminé par l'égalité (3) pour $p=k+1$. L'égalité $\frac{\overline{\beta}_{k+1}}{(k+1)!}=\beta_{k+1}$ s'établit également de la manière suivante ; on a

$$\frac{y(x_{n+1})-2y(x_n)+y(x_{n-1})}{h^2}=\sum_{p=0}^{k+1}\beta_p\nabla^p y''(x_n)+O(h^{k+2}).$$

Cette relation et l'égalité (4) entraînent

$$\frac{\overline{\beta}_{k+1}}{(k+1)!}\,y^{(k+3)}(\overline{x}_n)\,h^{k+1}=\beta_{k+1}\nabla^{k+1}y''(x_n)+O(h^{k+2});$$

en divisant les deux membres par $y^{(k+1)}(\overline{x}_n)\,h^{k+1}$ et en passant à la limite lorsque $h\to 0$ et pour $x_n=\text{const}$, on trouve

$$\frac{\overline{\beta}_{k+1}}{(k+1)!}=\beta_{k+1}.$$

On aboutit finalement à l'égalité

$$\frac{y(x_{n+1})-2y(x_n)+y(x_{n-1})}{h^2}=\sum_{p=0}^{k}\beta_p\nabla^p y''(x_n)+\beta_{k+1}y^{(k+3)}(\overline{x}_n)\,h^{k+1}$$

et à la formule de calcul correspondante

$$y_{n+1}-2y_n+y_{n-1}=h^2\sum_{p=0}^{k}\beta_p\nabla^p f_n,\quad f_m=f(x_m,\ y_m).$$

Si on remplace $y''(x)$ par un polynôme d'interpolation aux points x_{n+1}, ..., x_{n-k+1}, on obtient

$$\frac{y(x_{n+1})-2y(x_n)+y(x_{n-1})}{h^2}=\sum_{p=0}^{k}\gamma_p\nabla^p y''(x_{n+1})+\gamma_{k+1}y^{(k+3)}(\widetilde{x}_n)\,h^{k+1},$$

où

$$\gamma_p=\frac{1}{p!}\left(\int_0^1(1-t)(t-1)\,t\,\ldots\,(t+(p-2))\,dt+\right.$$

$$\left.+\int_0^{-1}(-1-t)(t-1)\,t\,\ldots\,(t+(p-2))\,dt\right),$$

et la formule d'interpolation correspondante

$$y_{n+1}-2y_n+y_{n-1}=h^2\sum_{p=0}^{k}\gamma_p\nabla^p f_{n+1}. \tag{5}$$

Citons plusieurs valeurs

$$\{\beta_p\}=1,\ 0,\ \frac{1}{12},\ \frac{1}{12},\ \frac{19}{240},\ \frac{3}{40},\ \ldots,$$

$$\{\gamma_p\}=1,\ -1,\ \frac{1}{12},\ 0,\ -\frac{1}{240},\ -\frac{1}{240},\ \ldots$$

Aux méthodes construites ci-dessus on donne parfois le nom de *méthodes d'Adams-Störmer.*

Afin d'intégrer numériquement l'équation

$$y''=f(x,\ y,\ y')$$

on se sert du procédé suivant: y_n et z_n, les valeurs approchées de $y(x_n)$ et $y'(x_n)$, se calculent par des relations de récurrence *explicites* telles que

$$\begin{cases} z_{n+1}=z_n+h\sum_{k=0}^{l}a_k\nabla^k f_n, \\ y_{n+1}=y_n+h\sum_{k=0}^{l}b_k\nabla^k z_{n+1} \end{cases}$$

ou *implicites* de la forme

$$\begin{cases} z_{n+1}=z_n+h\sum_{k=0}^{l}c_k\nabla^k f_{n+1}, \\ y_{n+1}=y_n+h\sum_{k=0}^{l}d_k\nabla^k z_{n+1}. \end{cases}$$

Les équations d'ordre supérieur à 2 se ramènent généralement à des systèmes d'équations d'ordre 1 ou 2.

§ 14. Evaluation de l'erreur dans la résolution numérique du problème de Cauchy pour une équation du deuxième ordre

Faisons comme dans le cas des schémas aux différences finies pour des équations du premier ordre et rejetons, dans (13.1) et (13.2), les termes en f. L'équation aux différences obtenue s'écrit dans les deux cas

$$y_{n+1} - 2y_n + y_{n-1} = 0.$$

Son équation caractéristique possède une racine multiple $\mu = 1$. Ceci ne doit pas nous effrayer : si dans les schémas aux différences approchant des équations du premier ordre les valeurs de f étaient multipliées par un coefficient en h, elles le seront par un coefficient en h^2.

Tous les schémas aux différences d'intégration de $y'' = f(x, y)$ que nous avons examinés peuvent se mettre sous la forme

$$\sum_{i=0}^{k} \frac{a_{-i} y_{n-i}}{h^2} - \sum_{i=0}^{k} b_{-i} f(x_{n-i}, y_{n-i}) = 0, \qquad a_0 \neq 0. \tag{1}$$

Soit r_n l'erreur d'approximation

$$r_n = \sum_{i=0}^{k} \frac{a_{-i} y(x_{n-i})}{h^2} - \sum_{i=0}^{k} b_{-i} f(x_{n-i}, y(x_{n-i})) \tag{2}$$

et supposons que $|r_n| \leqslant c_0 h^m$ uniformément sur tout le segment d'intégration considéré $[x_0, x_0 + X]$. Etant donné que nous arrondissons lors du calcul et que l'équation n'est résolue en y_n qu'avec une certaine précision, dans le cas d'un schéma implicite ($b_0 \neq 0$), la solution approchée y_n obtenue réellement vérifie généralement (1) avec une erreur. On a donc

$$\sum_{i=0}^{k} a_{-i} y_{n-i} - h^2 \sum_{i=0}^{k} b_{-i} f(x_{n-i}, y_{n-i}) = \delta_n.$$

Récrivons (2)

$$\sum_{i=0}^{k} a_{-i} y(x_{n-i}) - h^2 \sum_{i=0}^{k} b_{-i} f(x_{n-i}, y(x_{n-i})) = h^2 r_n ;$$

retranchant cette relation de la précédente nous trouvons l'équation d'erreur $R_n = y_n - y(x_n)$:

$$\sum_{i=0}^{k} a_{-i} R_{n-i} - h^2 \sum_{i=0}^{k} b_{-i} l_{n-i} R_{n-i} = g_n, \tag{3}$$

où

$$g_n = \delta_n - h^2 r_n, \qquad l_j = f_y(x_j, \tilde{y}_j).$$

Soit

$$L = \sup_{x_0 \leqslant x \leqslant x_0 + X} |f_y(x, y)| < \infty.$$

THÉORÈME. *Supposons que toutes les racines de l'équation caractéristique*

$$\sum_{i=0}^{k} a_{-i}\mu^{k-i} = 0 \tag{4}$$

d'un schéma aux différences sont situées dans le cercle unité et qu'il n'y a pas de racines multiples sur le contour du cercle, sauf une racine double 1. *Alors, pour* $x_0 \leqslant x_n \leqslant x_0 + X$, *la solution de l'équation* (3) *satisfait à*

$$\max_n \left(|R_n|, \left| \frac{R_n - R_{n-1}}{h} \right| \right) \leqslant$$

$$\leqslant C(L, X) \left(\frac{1}{h} \sum_{j=k}^{n} |g_j| + \max_{0 \leqslant j < k} |R_j| + \max_{0 < j < k} \left| \frac{R_j - R_{j-1}}{h} \right| \right). \tag{5}$$

DÉMONSTRATION. Posons $R_j - R_{j-1} = hz_j$. Selon l'hypothèse du lemme, l'équation caractéristique (4) admet une racine $\mu = 1$ et son premier membre s'écrit donc

$$\left(\sum_{i=0}^{k-1} d_{-i}\mu^{k-1-i} \right)(\mu - 1), \quad \text{où} \quad d_0 = a_0.$$

Soit E, comme plus haut, l'opérateur de translation $Ey_j = y_{j+1}$. On a la chaîne d'égalités

$$\sum_{i=0}^{k} a_{-i} R_{n-i} = \sum_{i=0}^{k} a_{-i} E^{k-i} R_{n-k} =$$

$$= \left(\sum_{i=0}^{k-1} d_{-i} E^{k-1-i} \right)(E-1) R_{n-k} = \sum_{i=0}^{k-1} d_{-i} E^{k-1-i} h z_{n-k+1};$$

et finalement

$$\sum_{i=0}^{k} a_{-i} R_{n-i} = h \sum_{i=0}^{k-1} d_{-i} z_{n-i}.$$

La relation (3) prend alors la forme

$$\sum_{i=0}^{k-1} d_{-i} z_{n-i} - h \sum_{i=0}^{k} b_{-i} l_{n-i} R_{n-i} = g_n h^{-1}.$$

Portons à droite tous les termes, sauf $d_0 z_n$, et divisons cette égalité par $d_0 \neq 0$; il vient

$$z_n = \sum_{i=1}^{k-1} \left(-\frac{d_{-i}}{d_0}\right) z_{n-i} + h \sum_{i=0}^{k} \left(\frac{b_{-i}}{d_0}\right) l_{n-i} R_{n-i} + \frac{g_n}{d_0 h}. \qquad (6)$$

Introduisons les vecteurs

$$\mathbf{u}_n = \begin{pmatrix} R_n \\ \cdots \\ R_{n-k+1} \end{pmatrix}, \qquad \mathbf{w}_n = \begin{pmatrix} z_n \\ \cdots \\ z_{n-k+2} \end{pmatrix}, \qquad \mathbf{v}_n = \begin{pmatrix} \mathbf{u}_n \\ \mathbf{w}_n \end{pmatrix}$$

et mettons les relations $R_n = R_{n-1} + h z_n$ et (6) sous la forme matricielle équivalente

$$\mathbf{v}_n = A\mathbf{v}_{n-1} + hD_n \mathbf{v}_n + hB_n \mathbf{v}_{n-1} + \mathbf{q}_n; \qquad (7)$$

les matrices et le vecteur $\mathbf{q}_n$ sont décomposés en blocs correspondant aux vecteurs $\mathbf{u}_n$ et $\mathbf{w}_n$:

$$A = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}, \quad D_n = \begin{pmatrix} D_{11} & D_{12} \\ D_{21}^n & D_{22} \end{pmatrix},$$

$$B_n = \begin{pmatrix} B_{11} & B_{12} \\ B_{21}^n & B_{22} \end{pmatrix}, \quad \mathbf{q}_n = \begin{pmatrix} \mathbf{q}_1^n \\ \mathbf{q}_2^n \end{pmatrix},$$

$$A_{11} = \begin{pmatrix} 1 & 0 & 0 & 0 & \ldots & 0 & 0 \\ 1 & 0 & 0 & 0 & \ldots & 0 & 0 \\ 0 & 1 & 0 & 0 & \ldots & 0 & 0 \\ 0 & 0 & 1 & 0 & \ldots & 0 & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & 0 & \ldots & 1 & 0 \end{pmatrix},$$

$$A_{22} = \begin{pmatrix} -\frac{d_{-1}}{d_0} & \cdots & -\frac{d_{2-k}}{d_0} & -\frac{d_{1-k}}{d_0} \\ 1 & 0\ 0\ 0\ \ldots & 0 & 0 \\ 0 & 1\ 0\ 0\ \ldots & 0 & 0 \\ 0 & 0\ 1\ 0\ \ldots & 0 & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0\ 0\ 0\ \ldots & 1 & 0 \end{pmatrix},$$

$$D_{12}=\begin{pmatrix} 1 & 0 & 0 & \dots & 0 \\ 0 & 0 & 0 & \dots & 0 \\ 0 & 0 & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & 0 \end{pmatrix}, \quad D_{21}^{n}=\begin{pmatrix} \frac{b_0 l_n}{d_0} & 0 & 0 & \dots & 0 \\ 0 & 0 & 0 & \dots & 0 \\ 0 & 0 & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & 0 \end{pmatrix},$$

$$B_{21}^{n}=\begin{pmatrix} \frac{b_{-1} l_{n-1}}{d_0} & \dots & \frac{b_{-k} l_{n-k}}{d_0} \\ 0 & \dots & 0 \\ \dots & \dots & \dots \\ 0 & \dots & 0 \end{pmatrix}, \quad \mathbf{q}_2^n=\begin{pmatrix} \frac{g_n}{d_0 h} \\ 0 \\ \dots \\ 0 \end{pmatrix}.$$

Les matrices A_{12}, A_{21}, D_{11}, D_{22}, B_{11}, B_{12}, B_{22} et les vecteurs $\mathbf{q}_1^n$ sont nuls. La matrice A est bloc-diagonale: $\begin{pmatrix} A_{11} & 0 \\ 0 & A_{22} \end{pmatrix}$. A_{11} a pour équation caractéristique $(\mu-1)\mu^{k-1}=0$, et A_{22} se présente comme la matrice A du § 8, de plus l'équation

$$\sum_{i=0}^{k-1} d_{-i}\mu^{k-1-i}=0$$

vérifie les conditions imposées à

$$\sum_{i=0}^{k} a_{-i}\mu^{k-i}=0$$

correspondant à cette dernière matrice. Il découle de ce que nous venons de dire que la matrice A jouit de la propriété suivante: son équation caractéristique ne possède pas de racines dont le module soit supérieur à 1 et quant aux racines de module 1, il leur correspond des boîtes de Jordan de dimension 1.

Portons l'expression de $hD_n\mathbf{v}_n$ dans le premier membre de (7) et multiplions à gauche par $(E-hD_n)^{-1}$ l'équation ainsi obtenue; il vient

$$\mathbf{v}_n=(E-hD_n)^{-1}(A+hB_n)\,\mathbf{v}_{n-1}+(E-hD_n)^{-1}\mathbf{q}_n.$$

Cette égalité peut s'écrire

$$\mathbf{v}_n=A\mathbf{v}_{n-1}+h(E-hD_n)^{-1}(D_nA+B_n)\,\mathbf{v}_{n-1}+(E-hD_n)^{-1}\mathbf{q}_n. \quad (8)$$

La relation (8) a une forme analogue à (8.7); la matrice A vérifie les conditions du § 8; si

$$\left|\frac{b_0 Lh}{d_0}\right|\leqslant\frac{1}{2}, \text{ alors } \|hD_n\|_1=\left|\frac{b_0 Lh}{d_0}\right|\leqslant\frac{1}{2},$$

et, selon (6.5.7), on a

$$\|(E-hD_n)^{-1}\|_1\leqslant 2.$$

D'où les inégalités

$$\| (E - hD_n)^{-1} (D_n A + B_n) \|_1 \leqslant \gamma_1 L,$$
$$\| (E - hD_n)^{-1} q_n \|_1 \leqslant 2 \| q_n \|_1.$$

L'égalité (8) satisfait donc aux mêmes conditions que (8.7). De la même façon dont on a déduit (8.12) à partir de (8.7), on obtient

$$\| v_n \|_1 \leqslant \| C \|_1 \exp(\gamma_1 LX) \left(\beta_1 \sum_{j=k}^{n} \| q_j \|_1 + \| C^{-1} \|_1 \| v_{k-1} \|_1 \right).$$

Puisque

$$\| q_j \|_1 = \left| \frac{g_j}{h d_0} \right|,$$
$$\| v_n \|_1 \geqslant \max \left(|R_n|, \left| \frac{R_n - R_{n-1}}{h} \right| \right),$$
$$\| v_{k-1} \|_1 = \max \left(\max_{0 \leqslant j < k} |R_j|, \max_{0 < j \leqslant k} \left| \frac{R_j - R_{j-1}}{h} \right| \right),$$

il en résulte notre théorème.

L'estimation

$$|g_j| \leqslant \max_j |\delta_j| + c_0 h^{m+2}$$

nous permet d'obtenir

$$\sum_{j=k}^{n} |g_j| \leqslant n \max_j |\delta_j| + c_0 n h^{m+2} \leqslant \frac{X}{h} \max_j |\delta_j| + c_0 X h^{m+1}.$$

En évaluant le second membre de (5) on a l'estimation d'erreur

$$\max \left(|R_n|, \frac{|R_n - R_{n-1}|}{h} \right) \leqslant C(L, X) \left(\frac{X}{h^2} \max_i |\dot{\delta}_j| + \right.$$
$$\left. + c_0 X h^m + \max_{0 \leqslant j < k} |R_j| + \max_{0 < j \leqslant k} \left| \frac{R_j - R_{j-1}}{h} \right| \right). \quad (9)$$

Dans le cas de l'équation $y' = My$ nous avons montré qu'il y a parmi les racines de

$$\sum_{i=0}^{k} (a_{-i} - Mhb_{-i}) \mu^{k-i} = 0$$

la racine $\mu = \exp(Mh) + O((Mh)^{m+1})$ si ce schéma approxime l'équation différentielle avec une erreur $O(h^m)$,

Dans le cas $y'' = My$, l'équation aux différences (1) s'écrit

$$\sum_{i=0}^{k} (a_{-i} - Mh^2 b_{-i}) y_{n-i} = 0.$$

On démontre de même que si le schéma (1) approche $y'=f(x, y)$ avec une erreur $O(h^m)$, parmi les racines de

$$\sum_{i=0}^{k}(a_{-i}-Mh^2b_{-i})\,\mu^{k-i}=0$$

il y a obligatoirement les deux racines

$$\mu_{1,2}=\exp(\pm\sqrt{M}h)+O((\sqrt{M}h)^{m+1}).$$

§ 15. Méthodes bilatérales

Le but des méthodes bilatérales est de localiser numériquement la solution exacte du problème proposé. Voyons quelques-unes d'entre elles en négligeant, pour la simplicité de l'exposé, l'influence de l'erreur de calcul.

Etant donnés $x_0<\ldots<x_n$ et les valeurs $f(x_0), \ldots, f(x_n)$, supposons que la dérivée $f^{(n)}(x)$ conserve son signe sur $[x_0, x_n]$. Proposons-nous de trouver l'intervalle contenant la valeur $f(x)$, $x\in(x_0, x_n)$. Posons

$$A_0=L_{0,\ \ldots,\ n-1}(x),\quad A_1=L_{1,\ \ldots,\ n}(x), \tag{1}$$

où $L_{k,\ \ldots,\ n-1+k}(x)$, $k=0, 1$, est un polynôme d'interpolation de degré $n-1$ basé sur $x_k, \ldots, x_{n-1+k}$. On a (2.3.1) la représentation du reste

$$f(x)-A_0=\frac{f^{(n)}(\zeta_0)}{n!}(x-x_0)\ldots(x-x_{n-1}),$$

$$f(x)-A_1=\frac{f^{(n)}(\zeta_1)}{n!}(x-x_1)\ldots(x-x_n),$$

$$\zeta_0,\ \zeta_1\subset[x_0, x_n].$$

Les facteurs $x-x_0$ et $x-x_n$ sont de signes contraires; il en découle que $f(x)-A_0$ et $f(x)-A_1$ présentent, elles aussi, des signes différents (si aucune d'elles ne s'annule). Aussi $f(x)$ est compris entre A_0 et A_1, autrement dit,

$$\min(A_0,\ A_1)\leqslant f(x)\leqslant\max(A_0,\ A_1).$$

Soit à calculer

$$I(f)=\int_0^1 f(x)\,dx.$$

Mettons cette intégrale sous la forme

$$I(f)=\sum_{q=1}^{M}I_q(f),\quad I_q(f)=\int_{\Delta_q}f(x)\,dx;\quad \Delta_q=\left[\frac{q-1}{M},\ \frac{q}{M}\right].$$

Trouvons $I_q(f)$ par la formule des trapèzes et celle des rectangles

$$I_q(f)\approx t_q(f)=\frac{1}{2M}\left(f\left(\frac{q-1}{M}\right)+f\left(\frac{q}{M}\right)\right),$$

$$I_q(f)\approx p_q(f)=\frac{1}{M}f\left(\frac{q-\frac{1}{2}}{M}\right).$$

Au § 1 du chap. III nous avons obtenu l'expression des erreurs sur ces formules :

$$I_q(f)-t_q(f)=-\frac{1}{12M^3}f''(x_q^0), \tag{2}$$

$$I_q(f)-p_q(f)=\frac{1}{24M^3}f''(x_q^1),$$

$$x_q^0,\ x_q^1\in\Delta_q.$$

Voyons d'abord le cas où $f''(x)$ ne change pas de signe (prenons $f''(x)\geqslant 0$ pour fixer les idées). Alors (2) entraîne

$$p_q(f)\leqslant I_q(f)\leqslant t_q(f),$$

et la sommation sur q donne

$$P_M(f)\leqslant I(f)\leqslant T_M(f),$$

avec

$$P_M(f)=\sum_{q=1}^{M}p_q(f),\quad T_M(f)=\sum_{q=1}^{M}t_q(f).$$

Si, par contre, $f''(x)$ change de signe sur $[0, 1]$, on utilise des stratégies différentes pour encadrer la valeur de l'intégrale.

En voici une, la moins sûre. La sommation de (2) sur q fournit

$$\begin{aligned} I(f)&=T_M(f)-\frac{1}{12M^2}S_M^0,\\ I(f)&=P_M(f)+\frac{1}{24M^2}S_M^1; \end{aligned} \tag{3}$$

ici

$$S_M^k=\frac{1}{M}\sum_{q=1}^{M}f''(x_q^k),\quad k=0,\ 1.$$

Soit $f''(x)$ bornée en module : $|f''(x)|\leqslant B<\infty$. Nous invitons le lecteur à prouver la validité des relations

$$S_M^k=\int_0^1 f''(x)\,dx+\varepsilon_M^k,\ |\varepsilon_M^k|\leqslant\frac{B}{2M},\quad k=0,\ 1. \tag{4}$$

Pour fixer les idées, considérons le cas

$$A=\int_0^1 f''(x)\,dx>0.$$

Alors

$$I(f)=T_M(f)-\frac{1}{12M^2}\left(\int_0^1 f''(x)\,dx+\varepsilon_M^0\right)\leqslant$$

$$\leqslant T_M(f)-\frac{1}{12M^2}\varepsilon_M^0\leqslant T_M(f)+\frac{B}{24M^3};$$

de même

$$I(f)=P_M(f)+\frac{1}{24M^2}\left(\int_0^1 f''(x)\,dx+\varepsilon_M^1\right)>P_M(f)-\frac{B}{48M^3}.$$

D'où l'estimation

$$P_M(f)-\frac{B}{48M^3}\leqslant I(f)\leqslant T_M(f)+\frac{B}{24M^3}. \tag{5}$$

Celle-ci n'est d'aucune utilité car seules les quantités $P_M(f)$ et $T_M(f)$ se déterminent numériquement. Avec un pas d'intégration si petit que $B/(2M)\leqslant A$, on a, conformément à (4),

$$\operatorname{sg} S_M^0=\operatorname{sg} S_M^1=\operatorname{sg} A.$$

Les relations (3) entraînent alors l'estimation effective

$$P_M(f)\leqslant I(f)\leqslant T_M(f).$$

Lorsque $A<0$, on a naturellement

$$T_M(f)\leqslant I(f)\leqslant P_M(f).$$

Regroupons ces inégalités; il vient

$$\min(P_M(f),\ T_M(f))\leqslant I(f)\leqslant\max(P_M(f),\ T_M(f)) \tag{6}$$

pour A de signe quelconque à condition que

$$\frac{B}{2M}\leqslant|A|.$$

Dans la pratique on estime que (6) est juste avec M quelconque bien que, par exemple pour $A=0$, cette affirmation soit gratuite.

Une autre stratégie plus prudente a formellement droit de cité pour $A=0$. Considérons ceux des segments Δ_q sur lesquels $f''(x)$ garde son signe. Lorsque $f''(x)\geqslant 0$, il découle de (2):

$$I_q(f)\leqslant t_q(f)-\frac{1}{12M^3}F_q^1,$$

$$I_q(f)\geqslant p_q(f)+\frac{1}{24M^3}F_q^1;$$

ici et dans la suite

$$F_q^1=\min_{\Delta_q}|f''(x)|,$$

$$F_q^2=\max_{\Delta_q}|f''(x)|.$$

De même, lorsque $f''(x)\leqslant 0$,

$$I_q(f)\geqslant t_q(f)+\frac{1}{12M^3}F_q^1,$$

$$I_q(f)\leqslant p_q(f)-\frac{1}{24M^3}F_q^1.$$

Ces relations entraînent que pour $f''(x)$ de signe constant sur Δ_q on a

$$s_q^1(f)+\frac{1}{24M^3}F_q^1\leqslant I_q(f)\leqslant s_q^2(f)-\frac{1}{24M^3}F_q^1, \tag{7}$$

avec

$$s_q^1(f) = \min(p_q(f),\ t_q(f)),$$
$$s_q^2(f) = \max(p_q(f),\ t_q(f)).$$

Supposons que $f''(x)$ change de signe en un point $\bar{x}_q$ de Δ_q. De tels segments sont évalués à l'aide de (2):

$$s_q^1(f) - \frac{F_q^2}{12M^3} \leqslant I_q(f) \leqslant s_q^2 + \frac{F_q^2}{12M^3}. \tag{8}$$

Désignons par $\sum'$ la sommation sur les segments avec $f''(x)$ d'un même signe et par $\sum''$ la sommation sur les segments restants. En sommant (7), (8) par rapport à tous les q nous trouvons

$$S_M^1(f) + E \leqslant I(f) \leqslant S_M^2(f) - E,$$

où

$$S_M^k(f) = \sum_{q=1}^{M} s_q^k(f), \quad k = 1,\ 2,$$

$$E = \sum_q{}' \frac{1}{24M^3} F_q^1 - \sum_q{}'' \frac{F_q^2}{12M^3}.$$

PROBLÈME. Démontrer

$$E > \frac{1}{24M^2} \int_0^1 |f''(x)|\,dx - \frac{B}{8M^3}.$$

Nous rappelons au lecteur que $\sum_q''$ contient des termes correspondant aux segments où $f''(x)$ change de signe; c'est pourquoi $F_q^2 = O\left(\frac{1}{M}\right)$ dans ces termes. D'une façon générale les points où la dérivée change de signe sont peu nombreux et $E > 0$ dès que M est relativement élevé. Ce qui autorise l'estimation

$$S_M^1(f) \leqslant I(f) \leqslant S_M^2(f). \tag{9}$$

On a les inégalités évidentes:

$$S_M^1(f) = \sum_{q=1}^{M} \min(p_q(f),\ t_q(f)) \leqslant$$
$$\leqslant \min\left(\sum_{q=1}^{M} p_q(f),\ \sum_{q=1}^{M} t_q(f)\right) = \min(P_M(f),\ T_M(f)),$$

de même,

$$\max(P_M(f),\ T_M(f)) \leqslant S_M^2(f).$$

Par conséquent, l'inégalité (9) donne de plus larges limites de variation de $I(f)$. Il est donc naturel qu'elle ait lieu avec des contraintes plus faibles sur le pas.

Passons maintenant aux méthodes bilatérales de résolution du problème de Cauchy pour l'équation différentielle $y' = f(x, y)$, $y(x_0) = y_0$. On désire

trouver la valeur $y(x_0 + X)$. Prenons un entier m arbitraire et faisons recours à deux méthodes d'Adams d'ordre d'exactitude m:

$$y_j - y_{j-1} = h \sum_{j=0}^{m} \gamma_i \nabla^i f_{j-1} \quad \text{(méthode explicite)}$$

et

$$y_j - y_{j-1} = h \sum_{j=0}^{m} \bar{\gamma}_i \nabla^i f_j \quad \text{(méthode implicite).}$$

On constate que les parties principales de l'erreur d'approximation dans ces méthodes, i.e.

$$\gamma_{m+1} y^{(m+2)}(x_j)\, h^{m+1} \quad \text{et} \quad \bar{\gamma}_{m+1} y^{(m+2)}(x_j)\, h^{m+1},$$

sont de signes opposés parce que $\gamma_{m+1}\bar{\gamma}_{m+1} < 0$.

La résolution d'un même problème de Cauchy par les deux techniques donne les valeurs approchées y_N^1 et y_N^2 de la solution $y(x_0 + X)$. Dans les conditions du théorème de la partie principale de l'erreur (§ 9 de ce chapitre) on a

$$\begin{aligned} y_N^1 - y(x_0 + X) &= -\gamma_{m+1} z(X)\, h^{m+1} + o(h^{m+1}), \\ y_N^2 - y(x_0 + X) &= -\bar{\gamma}_{m+1} z(X)\, h^{m+1} + o(h^{m+1}), \end{aligned} \tag{10}$$

avec

$$z(X) = \int_{x_0}^{x_0+X} y^{(m+1)}(t) \exp\left(\int_{t}^{x_0+X} f_y(\tau, y(\tau))\, d\tau \right) dt.$$

On en conclut que lorsque $z(X) \neq 0$ et h suffisamment petits,

$$y_N^1 - y(x_0 + X) \quad \text{et} \quad y_N^2 - y(x_0 + X)$$

sont de signes opposés, donc

$$\min(y_N^1, y_N^2) \leqslant y(x_0 + X) \leqslant \max(y_N^1, y_N^2). \tag{11}$$

Cette stratégie de la localisation de $y(x_0 + X)$ est équivalente au premier procédé d'encadrement de $I(f)$ qui est plus risqué. Néanmoins, l'inégalité (11) appelle une prudence plus grande que (6) du moment que dans le cas d'équations différentielles le terme d'ordre supérieur dans (10) ne devient souvent inférieur à la partie principale que si les pas sont très petits.

On ne connaît pas dans la pratique de variante rationnelle de la deuxième stratégie de localisation de la solution par les méthodes d'Adams.

Cette stratégie est à recommander pour les méthodes de Runge-Kutta. Bornons-nous, pour fixer les idées, aux méthodes exactes à l'ordre 2. Etant donné que les parties principales de l'erreur par pas dans les méthodes de Runge-Kutta d'un même ordre d'exactitude ne sont pas forcément proportionnelles, on est donc obligé à construire des procédés tels que ces parties principales soient proportionnelles et de signes différents.

Voici ce couple de procédés: dans les deux cas on calcule

$$k_1 = hf(x, y), \quad k_2 = hf\left(x + \frac{1}{3}h,\ y + \frac{1}{3}k_1\right),$$

puis

$$k_2^{(1)} = hf\left(x + \frac{1}{2}h,\ y + \frac{1}{2}k_2\right)$$

et

$$y(x+h) \approx y^{(1)} = y(x) + k_3^{(1)}$$

pour la première méthode, et

$$k_3^{(2)} = hf\left(x + \frac{5}{6}h,\ y + \frac{5}{6}k_2\right)$$

et

$$y(x+h) \approx y^{(2)} = y(x) + \frac{2}{5}k_1 + \frac{3}{5}k_3^{(2)}$$

pour la deuxième. L'erreur par pas dans la première méthode (cf. § 2, chap. VIII) s'écrit

$$\varphi(h) = y(x+h) - y^{(1)} = Ah^3 + O(h^4),$$

et celle dans la deuxième

$$\varphi(h) = y(x+h) - y^{(2)} = -Ah^3 + O(h^4),$$

où

$$A = \frac{f_{yy}f^2}{24}.$$

L'encadrement selon une stratégie plus prudente s'effectue comme suit. Pendant l'intégration on calcule à chaque pas deux quantités y_n^1 et y_n^2 encadrant la solution et telles que $y_n^1 < y_n^2$. On suppose $y_0^1 = y_0^2 = y(x_0)$. Avec y_n^1 au départ on effectue un pas d'après les deux méthodes et on prend pour y_{n+1}^1 la plus petite des valeurs $y_{n+1}^{1,1}$, $y_{n+1}^{1,2}$ ainsi obtenues. Ensuite, on part de y_n^2, on effectue un pas selon les deux méthodes et on prend pour y_{n+1}^2 la plus grande des valeurs $y_{n+1}^{2,1}$, $y_{n+1}^{2,2}$ trouvées.

Si (ce qui arrive fort rarement dans la pratique) $y_n^1 > y_n^2$ pour un certain n, cette méthode bilatérale tombe en défaut.

Comme les points y_n^1 et y_n^2 sont assez voisins, les valeurs A qui leur correspondent, le sont elles aussi, et lorsque $A \neq 0$, on a généralement

$$y_{n+1}^2 = y_{n+1}^{2,2} \text{ si } y_{n+1}^1 = y_{n+1}^{1,1}$$

et (12)

$$y_{n+1}^2 = y_{n+1}^{2,1} \text{ si } y_{n+1}^1 = y_{n+1}^{1,2}.$$

Notons que, les méthodes examinées étant exactes à l'ordre 2, $y_n^1 - y(x_n)$ et $y_n^2 - y(x_n)$ sont en h^2.

Cependant, les parties principales de l'erreur par pas dans ces méthodes étant égales en valeur absolue et de signe différent, on a

$$\left| y_n - \frac{y_n^1 + y_n^2}{2} \right| = O(h^3).$$

Si les méthodes bilatérales permettent d'établir avec une certitude plus grande les limites de variation de la solution d'un problème, elles alourdissent par

contre l'algorithme et font augmenter le volume de calcul. Dans la pratique, les techniques d'encadrement décrites peuvent donner lieu à une erreur de calcul qui, quand on en tient compte, complique toujours plus les algorithmes en question et en limite extrêmement l'usage.

Bibliographie du chapitre VIII

I [1, 4, 11, 17], II [1, 31, 35]; §§ 1-3: III [54, 58]; § 4: III [9, 49]; §§ 5, 6: III [57, 59]; § 7: I [1], III [7, 55]; § 8: I [1], III [32]; § 9: II [27], III [6, 15, 48]; § 10: I [17], III [6, 15]; § 11: III [17, 37, 39*), 56, 60, 62]; § 12: I [5], III [6, 18, 56]; § 14: II [35]; § 15: III [16, 23, 41].

---

*) Voir également Y. Rakitski, « Nouvelles méthodes numériques de résolution des systèmes d'équations différentielles ordinaires et d'équations aux différences », Travaux de IPL, 332 (1973), 88-97 (en russe).

CHAPITRE IX

MÉTHODES DE RÉSOLUTION DES PROBLÈMES AUX LIMITES POUR LES ÉQUATIONS DIFFÉRENTIELLES ORDINAIRES

Par comparaison au problème de Cauchy, la résolution des problèmes aux limites présente des difficultés complémentaires: l'existence de la solution nécessite une étude notablement plus compliquée; la formulation du problème discrétisé donne lieu à un système d'équations linéaires ou non linéaires dont la résolution exige une étude spéciale.

§ 1. Méthodes simples de résolution du problème aux limites pour une équation du deuxième ordre

Une partie essentielle des problèmes aux limites pour les équations différentielles ordinaires est constituée par les problèmes relatifs aux équations et systèmes d'ordre 2, que l'on rencontre en balistique, en théorie de l'élasticité, etc.

Commençons notre étude par un problème particulier, assez fréquent. On recherche la solution de l'équation

$$Ly = y'' - p(x)\, y = f(x) \text{ sur } (0, X) \tag{1}$$

avec les conditions aux limites

$$y(0) = a, \quad y(X) = b. \tag{2}$$

Donnons-nous un pas $h = XN^{-1}$, N entier, et prenons les points $x_j = jh$ pour nœuds du réseau. Soit, comme d'ordinaire, y_j des approximations de $y(x_j)$. En remplaçant la dérivée $y''(x_j)$ par le rapport aux différences

$$\frac{\delta^2 y_j}{h^2} = \frac{y_{j+1} - 2y_j + y_{j-1}}{h^2},$$

il vient

$$l(y_j) = \frac{y_{j+1} - 2y_j + y_{j-1}}{h^2} - p_j y_j = f_j, \qquad j = 1, \ldots, N-1, \tag{3}$$

avec $p_j = p(x_j)$, $f_j = f(x_j)$; on substitue aux conditions aux limites les relations

$$y_0 = a, \quad y_N = b. \tag{4}$$

Montrons que, pour $p(x) \geqslant 0$, le système d'équations (3), (4) admet une solution et évaluons l'erreur.

LEMME 1. *Soit* $p(x) \geqslant 0$, $l(z_j) \leqslant 0$, z_0, $z_N \geqslant 0$. *Alors* $z_j \geqslant 0$ *quel que soit* j.

DÉMONSTRATION. Désignons $\min\limits_{0 \leqslant j \leqslant N} z_j$ par d. Supposons que $d < 0$, donc $d \neq z_0$, z_N. Soit q le plus petit entier tel que $z_q = d$; selon la définition de d et q, on a $z_{q-1} > d$, $z_{q+1} \geqslant d$. Alors

$$l(z_q) = \frac{(z_{q+1} - d) + (z_{q-1} - d)}{h^2} - p_q z_q \geqslant \frac{z_{q-1} - d}{h^2} > 0,$$

ce qui est contraire à l'hypothèse $d < 0$.

LEMME 2. *Si* $p(x) \geqslant 0$, *alors toute fonction* z_j *vérifie l'inégalité*

$$\max_{0 \leqslant j \leqslant N} |z_j| \leqslant \max(|z_0|, |z_N|) + Z \frac{X^2}{8},$$

où

$$Z = \max_{0 < j < N} |l(z_j)|.$$

DÉMONSTRATION. Faisons intervenir la fonction

$$w_j = |z_0| \left(1 - \frac{jh}{X}\right) + |z_N| \frac{jh}{X} + Z \frac{jh(X - jh)}{2}.$$

Pour un polynôme du deuxième degré, la quantité $\frac{\delta^2 f}{h^2}$ coïncide avec la dérivée seconde, aussi $\frac{\delta^2 w_j}{h^2} = -Z$. Puisque la définition de w_j entraîne $w_j \geqslant 0$, on a

$$l(w_j) = \frac{\delta^2 w_j}{h^2} - p_j w_j \leqslant -Z, \quad l(w_j \pm z_j) \leqslant -Z \pm l(z_j) \leqslant 0.$$

Evidemment,

$$w_0 \pm z_0 = |z_0| \pm z_0 \geqslant 0, \quad w_N \pm z_N = |z_N| \pm z_N \geqslant 0.$$

Les fonctions $w_j \pm z_j$ vérifient les conditions du lemme 1, c'est pourquoi elles sont supérieures ou égales à zéro. D'où l'estimation $|z_j| \leqslant |w_j| \leqslant \max\limits_{0 \leqslant j \leqslant N} |w_j|$. On a

$$|z_0| \left(1 - \frac{jh}{X}\right) + |z_N| \frac{jh}{X} \leqslant \max(|z_0|, |z_N|) \left(\left|1 - \frac{jh}{X}\right| + \left|\frac{jh}{X}\right| \right) =$$

$$= \max(|z_0|, |z_N|),$$

$$jh(X - jh) \leqslant \frac{X^2}{4}.$$

Aussi

$$\max_{0\leqslant j\leqslant N}|w_j|\leqslant\max(|z_0|,\ |z_N|)+Z\frac{X^2}{8}.$$

Le lemme 2 est démontré. Il en résulte que le système homogène correspondant à (3)-(4) n'admet que la solution triviale. Puisque le système non homogène (3)-(4) a autant d'équations que d'inconnues, il possède une solution et une seule.

Considérons le cas de $p(x)$ et $f(x)$ deux fois continûment dérivables; on démontre, dans les cours d'équations différentielles, que la solution $y(x)$ est alors quatre fois continûment dérivable.

Soit r_j l'erreur d'approximation correspondant au schéma aux différences finies (3):

$$r_j=l(y(x_j))-f_j=$$
$$=\frac{y(x_{j+1})-2y(x_j)+y(x_{j-1})}{h^2}-p(x_j)\,y(x_j)-f(x_j). \tag{5}$$

Comme $p(x)\,y(x)+f(x)=y''(x)$, on a

$$r_j=\frac{y(x_{j+1})-2y(x_j)+y(x_{j-1})}{h^2}-y''(x_j).$$

Les évaluations de l'erreur dans les formules de dérivation numérique donnent

$$r_j=\frac{y^{(4)}(\bar{x}_j)\,h^2}{12},\quad \text{où}\quad x_{j-1}\leqslant\bar{x}_j\leqslant x_{j+1}. \tag{6}$$

Vu divers arrondis, les approximations y_j des valeurs $y(x_j)$, qu'on obtient dans les calculs, vérifient le système (3), (4) avec les erreurs

$$l(y_j)-f_j=\delta_j. \tag{7}$$

En retranchant (5) de (7) on trouve l'équation

$$l(R_j)=\delta_j-r_j$$

par rapport à l'erreur $R_j=y_j-y(x_j)$ sur la solution approchée. Utilisons le lemme 2; on obtient

$$|R_j|\leqslant\max(|R_0|,\ |R_N|)+\frac{X^2}{8}\Big(\max_{0<j<N}|r_j|+\max_{0<j<N}|\delta_j|\Big).$$

Selon l'estimation (6)

$$|r_j|\leqslant M_4\frac{h^2}{12},\quad M_4=\max_{[0,\,X]}|y^{(4)}(x)|$$

et l'erreur est finalement appréciée par

$$\max_{0\leqslant j\leqslant N}|R_j|\leqslant\max(|R_0|,\ |R_N|)+\frac{M_4X^2h^2}{96}+\Big(\max_{0<j<N}|\delta_j|\Big)\frac{X^2}{8}.$$

Nous voyons que dès que les conditions aux limites et l'équation aux différences sont satisfaites avec une précision améliorée, pour un pas tendant vers zéro la solution du problème discrétisé se rapproche de celle du problème différentiel.

La méthode décrite donne une solution approchée qui converge vers la solution exacte avec la vitesse $O(h^2)$. Proposons-nous de construire des schémas plus précis et supposons que $p(x)$ et $f(x)$ sont quatre fois continûment dérivables; la solution du problème est alors 6 fois continûment dérivable. Revenons à la quantité

$$r_j = \frac{y(x_{j+1}) - 2y(x_j) + y(x_{j-1})}{h^2} - y''(x_j);$$

en développant toutes les quantités en série de Taylor au point x_j, il vient

$$r_j = \frac{y^{(4)}(x_j)\, h^2}{12} + O(h^4). \tag{8}$$

Retranchons de $l(y(x_j))$ un terme approximant $\frac{y^4(x_j)\, h^2}{12}$; il correspond au schéma obtenu une erreur d'approximation d'ordre plus élevé. On peut, par exemple, approcher $y^{(4)}(x_j)$ par

$$\frac{\delta^4 y(x_j)}{h^4} = \frac{y(x_{j+2}) - 4y(x_{j+1}) + 6y(x_j) - 4y(x_{j-1}) + y(x_{j-2})}{h^4}$$

et obtenir ainsi le schéma aux différences finies

$$\frac{\delta^2 y_j}{h^2} - \frac{\delta^4 y_j}{12h^2} - p_j y_j = f_j. \tag{9}$$

On arrive au même résultat en substituant directement à $y''(x_j)$ l'expression $h^{-2}(\delta^2 y(x_j) - (1/12)\delta^4 y(x_j))$ qui l'approche avec une erreur $O(h^4)$. L'équation (9) contient 5 inconnues y_j avec des coefficients non nuls; la résolution d'un système formé d'équations (9) et d'équations obtenues par approximation des conditions aux limites est plus laborieuse que celle du système (3).

En nous servant d'autres considérations, construisons un schéma aux différences finies où l'erreur d'approximation est $O(h^4)$ et tel que dans chaque équation figurent en tout trois inconnues. La double dérivation de l'équation de départ donne $y^{(4)}(x) = (p(x)\, y + f)''$. Aussi

$$y^{(4)}(x_j) \approx \frac{\delta^2 (py+f)\,|_{x_j}}{h^2} =$$

$$= \frac{(p_{j+1} y(x_{j+1}) + f_{j+1}) - 2(p_j y(x_j) + f_j) + (p_{j-1} y(x_{j-1}) + f_{j-1})}{h^2}.$$

La soustraction du schéma initial de l'approximation de $y^{(4)}h^2/12$ donne le schéma $\frac{\delta^2 y_j}{h^2} - p_j y_j - \frac{1}{12}\delta^2(p_j y_j + f_j) = f_j$ ou

$$l^{(1)}(y_j) = \frac{\delta^2 y_j}{h^2} - p_j y_j - \frac{1}{12}\delta^2(p_j y_j) = \overline{l}^{(1)}(f_j) = f_j + \frac{1}{12}\delta^2 f_j.$$

Dans l'hypothèse que la solution est 8 fois continûment dérivable, considérons l'erreur commise en approximant le nouveau schéma $r_j^{(1)} = l^{(1)}(y(x_j)) - \overline{l}^{(1)}(f_j)$, compte tenu de $p_j y(x_j) + f_j = y''(x_j)$. Il vient

$$r_j^{(1)} = l^{(1)}(y(x_j)) - \overline{l}^{(1)}(f_j) = \frac{\delta^2 y(x_j)}{h^2} - y''(x_j) - \frac{1}{12}\delta^2 y''(x_j).$$

Le développement de toutes les quantités en série de Taylor suivant les puissances de h au point x_j fournit

$$r_j^{(1)} = -\frac{y^{(6)}(x_j)\,h^4}{240} + O(h^6).$$

Construisons une approximation de $y^{(6)}(x_j)$ par $y(x_{j-1})$, $y(x_j)$, $y(x_{j+1})$. Ces approximations peuvent être très nombreuses. Procédons par exemple comme suit. Selon (1),

$$y'' = py + f, \quad y''' = (py+f)', \quad y^{(4)} = (py+f)'',$$
$$y^{(6)} = (py+f)^{(4)} = py^{(4)} + 4p'y^{(3)} + 6p''y'' + 4p^{(3)}y' + p^{(4)}y + f^{(4)},$$

et on a par conséquent l'égalité approchée

$$y^{(6)}(x_j) \approx p_j \frac{\delta^2(p_j y(x_j) + f_j)}{h^2} +$$
$$+ 4p'(x_j)\frac{(p_{j+1}y(x_{j+1}) + f_{j+1}) - (p_{j-1}y(x_{j-1}) + f_{j-1})}{2h} +$$
$$+ 6p''(x_j)(p_j y(x_j) + f_j) + 4p^{(3)}(x_j)\frac{y(x_{j+1}) - y(x_{j-1})}{2h} +$$
$$+ p^{(4)}(x_j)\,y(x_j) + f^{(4)}(x_j).$$

En ajoutant à $l^1(y_j)$ l'expression approchant $\frac{y^{(6)}(x_j)h^4}{240}$ on aboutit à un schéma aux différences finies avec une erreur d'approximation $O(h^6)$; chaque équation du système algébrique ainsi obtenu ne renferme que trois inconnues y_j.

L'évaluation réelle de l'erreur sur la solution d'un problème aux limites peut se faire par la règle de Runge dont la légitimité se base sur l'existence de la partie principale de l'erreur.

PROBLÈME. Soient $p(x)$ et $f(x)$ des fonctions 4 fois dérivables. Démontrer que la solution du problème (3)-(4) satisfait à

$$\max_n |y_n - y(x_n) - h^2 z(x_n)| = O(h^4),$$

où $z(x)$ est solution du problème aux limites

$$Lz = -y^4(x)/12, \quad z(0) = z(X) = 0.$$

INDICATION. Se servir des constructions du § 5, chap. VIII.

Passons à un autre groupe de méthodes d'amélioration de la précision des solutions approchées que nous appellerons *méthodes des différences d'ordre supérieur.*

Supposons qu'en résolvant le problème aux différences (3) on arrive à calculer les valeurs d'une quantité r_j^1 telle que

$$\max_{0<j<N} |\varepsilon_j^1| = O(h^4), \quad \varepsilon_j^1 = r_j - r_j^1.$$

Trouvons la solution R_j^1 du problème discret

$$l(R_j^1) = r_j^1, \quad j = 1, \ldots, N-1, \quad R_0^1 = R_N^1 = 0,$$

et posons $y_j^1 = y_j + R_j^1$. On a les égalités

$$l(y(x_j) - y_j^1) = l(y(x_j)) - l(y_j) - l(R_j^1) = r_j - r_j^1 = \varepsilon_j^1.$$

Compte tenu du lemme 2, il en résulte l'estimation

$$|y(x_j) - y_j^1| \leqslant \frac{X^2}{8} \max_{0<j<N} |\varepsilon_j^1| = O(h^4). \tag{10}$$

Conformément à (8) les quantités r_j^1 cherchées doivent vérifier

$$\max_j \left| r_j^1 - \frac{y^{(4)}(x_j) h^2}{12} \right| = O(h^4). \tag{11}$$

Si $p(x)$ et $f(x)$ sont suffisamment régulières, on approche $y^{(4)}(x_j)$ par $\frac{\delta^4 y_j}{h^4}$ et, partant, on pose $r_j^1 = \frac{\delta^4 y_j}{12h^2}$. Le terme à droite de cette relation n'est pas défini pour $j = 1$ et $j = N - 1$. Aussi, pour déterminer les approximations de $y^{(4)}(x_1)$ et $y^{(4)}(x_{N-1})$, utilise-t-on d'autres formules de dérivation numérique exactes à $O(h^4)$.

Notre affirmation sur la possibilité d'approcher les dérivées $y^{(4)}(x_j)$ exige une justification spéciale.

r_j^1 s'obtiennent également de la façon suivante. Vu (1), on a

$$y^{(4)}(x) = (y''(x))'' = (p(x)\, y(x) + f(x))'' = \frac{\delta^2 (p_j y(x_j) + f(x_j))}{h^2} + O(h^2),$$

ce qui suggère de poser

$$r_j^1 = \frac{\delta^2 (p_j y_j + f_j)}{12};$$

on montre que l'estimation (11) et donc (10) sont alors valables. Dans la pratique on se sert plus de la seconde technique d'obtention de r_j^1.

Le schéma général d'amélioration de la solution par différences d'ordre supérieur se présente comme suit.

On forme une équation qui vérifie approximativement l'erreur entachant l'approximation obtenue, puis on résout l'équation par rapport à cette erreur. Autrement dit, soit y_j^m une approximation avec une erreur $O(h^{2m+2})$.

On construit une expression de r_j^{m+1} qui se calcule efficacement au moyen des valeurs de y_j^m et telle que si R_j^{m+1} est solution du problème discrétisé

$$l(R_j^{m+1}) = r_j^{m+1}, \quad j = 1, \ldots, N-1, \quad R_0^{m+1} = R_N^{m+1} = 0,$$

l'approximation $y_j^{m+1} = y_j^m + R_j^{m+1}$ est affectée de l'erreur $O(h^{2m+1})$. r_j^{m+1}, R_j^{m+1}, puis y_j^{m+1} se calculent d'après le schéma ci-dessus.

Les méthodes décrites possèdent l'avantage de résoudre, quel que soit l'ordre de précision de l'approximation obtenue, les systèmes d'équations algébriques linéaires ayant une même matrice.

Traitons le cas de la condition aux limites $y'(0) - \alpha y(0) = a$. L'approximation discrète très précise d'une telle condition s'obtient directement en remplaçant la dérivée $y'(0)$ par une formule de dérivation numérique d'ordre de précision élevé

$$y'(0) \approx \sum_{i=0}^{l} \frac{c_i y(x_i)}{h}.$$

L'erreur d'approximation sera moindre et la résolution du système algébrique obtenu moins ardue si l'on améliore progressivement, comme ci-dessus, la précision de l'approximation.

La substitution de $\frac{y(h)-y(0)}{h}$ à $y'(0)$ donne l'équation $l_0^{(1)}(y_j) = \frac{y_1 - y_0}{h} - \alpha y_0 - a = 0$. Développons $r_0^{(1)} = l_0^{(1)}(y(x_j)) = \frac{y(h)-y(0)}{h} - \alpha y(0) - a$ en série de Taylor suivant les puissances de h; il vient

$$r_0^{(1)} = y'(0) + \frac{y''(0)h}{2} + O(h^2) - \alpha y(0) - a = \frac{y''(0)h}{2} + O(h^2).$$

L'erreur résultant de l'approximation de la condition aux limites est donc $O(h)$. Puisque, selon (1),

$$y''(0) = p_0 y(0) + f_0,$$

il correspond à

$$l_0^{(2)}(y_j) = \frac{y_1 - y_0}{h} - \alpha y_0 - a - (p_0 y_0 + f_0)\frac{h}{2} = 0$$

une approximation d'ordre 2. Le développement taylorien suivant les puissances de h de $r_0^{(2)} = l_0^{(2)}(y(x_j))$ fournit $r_0^{(2)} = \frac{y^{(3)}(0)h^2}{6} + O(h^3)$. On obtient en dérivant l'équation initiale (1):

$$y^{(3)}(0) - p_0 y'(0) - p'(0)\, y(0) - f'(0) = 0;$$

et, compte tenu de la condition aux limites,

$$y^{(3)}(0) = p_0(\alpha y(0) + a) + p'(0)\, y(0) + f'(0).$$

Il correspondra à l'équation aux différences

$$l_0^{(3)}(y_j) = l_0^{(2)}(y_j) - ((p_0\alpha + p'(0))\, y_0 + p_0 a + f'(0))\frac{h^2}{6} = 0$$

une approximation d'ordre 3.

Nous aurions pu écrire d'emblée

$$r_0^{(1)} = y''(0)\frac{h}{2} + y^{(3)}(0)\frac{h^2}{6},$$

puis exprimer les dérivées $y''(0)$ et $y^{(3)}(0)$ à l'aide de $y(0)$ et, en retranchant du schéma aux différences l'expression correspondante, obtenir $l_0^{(3)}(y_j) = 0$. Or, nous avons centré l'attention sur l'amélioration de proche en proche de la précision puisque l'application de cette technique à des équations aux dérivées partielles est la plus simple et la plus naturelle.

§ 2. Fonction de Green d'un problème aux limites discrétisé

La fonction w_j introduite au § 1 dans la démonstration du lemme 2 s'appelle *fonction majorante* et l'évaluation d'erreur faisant appel à cette fonction porte le nom de *méthode des majorantes* ou *de Gerschgorin*.

Dans certains cas où la méthode des majorantes ne joue pas, l'erreur sur la solution approchée est évaluée à l'aide de *l'analogue discret de la fonction de Green*. La construction de la fonction de Green du problème aux limites discrétisé (1.3)-(1.4) est en outre intéressante par son analogie avec le cas des problèmes aux limites différentiels.

La fonction de Green $G(x, s)$ du problème aux limites différentiel

$$Ly = y'' - p(x)\, y = f(x), \quad y(0) = a, \quad y(X) = b$$

est définie comme solution de l'équation

$$L(G(x, s)) = G''_{xx}(x, s) - p(x)\, G(x, s) = \delta(x - s) \qquad (1)$$

avec la condition aux limites $G(0, s) = G(X, s) = 0$. La solution du problème (1.1)-(1.2) se récrit à l'aide de la fonction de Green :

$$y(x) = \int_0^X G(x, s)\, f(s)\, ds + G_s(x, X)\, b - G_s(x, 0)\, a. \qquad (2)$$

Cette fonction peut être donnée par les formules explicites ci-dessous. Soient $W^1(x)$, $W^2(x)$ les solutions de l'équation $L(W) = 0$ vérifiant les conditions

$$W^1(0) = 0, \quad (W^1)'(0) = 1,$$
$$W^2(X) = 0, \quad (W^2)'(X) = -1 :$$

alors

$$G(x, s)=\begin{cases}\dfrac{W^2(s)\,W^1(x)}{V^0} & \text{pour } 0\leqslant x\leqslant s,\\ \dfrac{W^1(s)\,W^2(x)}{V^0} & \text{pour } s\leqslant x\leqslant X,\end{cases}$$

où

$$V(x)=\begin{vmatrix} W^1(x) & W^2(x)\\ (W^1)'(x) & (W^2)'(x)\end{vmatrix}=\text{const}=V^0.$$

Passons au problème discret

$$l(y_j)=\frac{y_{j+1}-2y_j+y_{j-1}}{h^2}-p_jy_j=f_j,\quad y_0=a,\quad y_N=b.$$

Définissons par analogie les fonctions W_n^1, W_n^2 moyennant les relations

$$l(W_j^i)=0,\qquad i=1,\ 2,\quad j=1,\ \ldots,\quad N-1,$$
$$W_0^1=0,\quad W_1^1=h,\quad W_N^2=0,\quad W_{N-1}^2=h.$$

L'analogue du wronskien est

$$V_n=\begin{vmatrix} W_n^1 & W_n^2\\ \dfrac{W_n^1-W_{n-1}^1}{h} & \dfrac{W_n^2-W_{n-1}^2}{h}\end{vmatrix}=\frac{W_n^2W_{n-1}^1-W_n^1W_{n-1}^2}{h}.$$

L'identité

$$0=W_n^1l(W_n^2)-W_n^2l(W_n^1)=$$
$$=\frac{W_n^1W_{n+1}^2+W_n^1W_{n-1}^2-W_n^2W_{n+1}^1-W_n^2W_{n-1}^1}{h^2}=\frac{V_{n+1}-V_n}{h^2}$$

entraîne que V_n est indépendant de n; notons-le $V(h)$.

Posons

$$G_n^k=\begin{cases}\dfrac{W_k^2W_n^1}{V(h)} & \text{pour } 0\leqslant n\leqslant k,\\ \dfrac{W_k^1W_n^2}{V(h)} & \text{pour } k\leqslant n\leqslant N\end{cases}$$

Introduisons la notation

$$l(G_n^k)=\frac{G_{n+1}^k-2G_n^k+G_{n-1}^k}{h^2}-p_nG_n^k.$$

La définition de W_n^i entraîne

$$l(G_n^k)=\frac{W_k^2}{V(h)}\,l(W_n^1)=0$$

lorsque $n < k$,

$$l(G_n^k) = \frac{W_k^1}{V(h)} l(W_n^2) = 0$$

lorsque $n > k$, et

$$l(G_n^k) = \frac{G_{n+1}^n - 2G_n^n + G_{n-1}^n}{h^2} - p_n G_n^n =$$

$$= \left(\frac{W_{n+1}^2 W_n^1 - 2W_h^1 W_n^2 + W_{n-1}^1 W_n^2}{V(h)\, h^2} - \frac{p_n W_n^1 W_n^2}{V(h)} \right) =$$

$$= \frac{W_n^1}{V(h)} \left(\frac{W_{n+1}^2 - 2W_n^2 + W_{n-1}^2}{h^2} - p_n W_n^2 \right) + \frac{(W_n^2 W_{n-1}^1 - W_h^1 W_{n-1}^2)}{V(h)\, h^2}$$

pour $k = n$. D'après la définition de W_n^2 la première parenthèse vaut 0 et la deuxième $V(h)\, h$. Finalement, pour $k = n$, on a $l(G_n^k) = h^{-1}$; en réunissant les relations obtenues, il vient

$$l(G_n^k) = \delta_{n-k} h^{-1}. \tag{3}$$

La fonction $\delta_{n-k} h^{-1}$ est l'analogue discret de la fonction de Dirac et l'égalité (3) celui de (1).

Evaluons le degré de proximité des fonctions de Green des problèmes différentiel et discret. Comme $(W^1)''_{xx}(0) = p(0)\, W^1(0) = 0$, on a

$$W^1(h) = W^1(0) + (W^1)'(0)\, h + \frac{(W^1)''(0)\, h^2}{2} + O(h^3) = h + O(h^3).$$

La fonction W_n^1 est solution de l'équation aux différences finies qui approche $W'' - p(x)\, W = 0$. L'équation caractéristique $\mu^2 - 2\mu + 1 = 0$ du schéma aux différences possède une racine double $\mu = 1$. Ainsi, les conditions du théorème du § 14, chap. VIII sont réalisées. Posons $R_n = W_n^1 - W^1(nh)$. Compte tenu de l'estimation (8.14.5) on a, sous les hypothèses concrètes du problème considéré,

$$\max_{0 \leqslant x_n \leqslant X} |R_n| \leqslant C(P, X) \left(h \sum_{i=1}^{N-1} |r_i| + \max(|R_0|, |R_1|) + \left| \frac{R_1 - R_0}{h} \right| \right),$$

où r_i est l'erreur d'approximation :

$$r_i = \frac{(W^1)^{(4)}(\bar{x}_i)\, h^2}{12}, \quad P = \max_{[0, X]} |p(x)|.$$

Par définition de W_0^1 et W_1^1, on a $R_0 = 0$, $R_1 = O(h^3)$, donc $\frac{R_1 - R_0}{h} = O(h^2)$. Il résulte par conséquent de la dernière évaluation

$$\max_{0 \leqslant x_n \leqslant X} |W_n^1 - W^1(nh)| \leqslant Mh^2 \tag{4}$$

avec M une constante. On obtient de même

$$\max_{0\leqslant x_n\leqslant X} |W_n^2 - W^2(nh)| \leqslant Mh^2. \tag{5}$$

Les propriétés de V^0 et $V(h)$ font que

$$V^0 = V(X) = \begin{vmatrix} W^1(X) & 0 \\ (W^1)'(X) & -1 \end{vmatrix} = -W^1(X),$$

$$V(h) = \begin{vmatrix} W_N^1 & 0 \\ \dfrac{W_N^1 - W_{N-1}^1}{h} & -1 \end{vmatrix} = -W_N^1.$$

C'est pourquoi

$$|V(h) - V^0| \leqslant Mh^2$$

conformément à l'*estimation* (4).

On suppose ensuite $W^1(X) \neq 0$; dans le cas contraire le problème aux limites homogène $y'' - p(x)y = 0$, $y(0) = y(X) = 0$ admet la solution $W^1(x)$ différente de zéro, et, par conséquent, le problème non homogène (1.1)-(1.2) ou bien n'a pas de solution, ou bien en a plus d'une. Supposons de plus h tellement petit que $2Mh^2 \leqslant |W^1(X)| = |V^0|$. On a alors

$$|V(h)| \geqslant |V^0| - Mh^2 \geqslant \left|\frac{V^0}{2}\right| > 0.$$

Pour de tels h

$$\left|\frac{1}{V(h)} - \frac{1}{V^0}\right| = \left|\frac{V^0 - V(h)}{V(h)\,V^0}\right| \leqslant \frac{2Mh^2}{|V^0|^2}. \tag{6}$$

La comparaison des expressions explicites des fonctions de Green des problèmes différentiel et discret donne, compte tenu des estimations (4)-(6),

$$|G_n^k - G(nh, kh)| \leqslant Qh^2 \tag{7}$$

lorsque $0 \leqslant kh$, $nh \leqslant X$.

LEMME. *Si* $V(h) \neq 0$, *le problème discrétisé* (1.3)-(1.4) *admet une solution unique qui s'écrit*

$$y_n = h\sum_{k=1}^{N-1} G_n^k f_k + \frac{W_n^2}{W_0^2}a + \frac{W_n^1}{W_N^1}b, \tag{8}$$

La formule (8) est l'analogue discret de la formule (2).

DÉMONSTRATION. Puisque $G_0^k = G_N^k = 0$ d'après la définition de G_n^k, $y_0 = a$, $y_N = b$. On a l'égalité

$$l(y_n) = h\sum_{k=1}^{N-1} l(G_n^k) f_k + \frac{l(W_n^2)}{W_0^2}a + \frac{l(W_n^1)}{W_N^1}b,$$

où les coefficients de a et b sont nuls par définition de W_n^i. En utilisant (3) on trouve $l(y_n) = f_n$. Ainsi, le système d'équations linéaires (1.3)-(1.4) possède, quels que soient ses deuxièmes membres, une solution sous forme (8). Cette solution est unique, comme c'est la règle pour tout système d'équations linéaires résoluble pour n'importe quels deuxièmes membres. Le lemme est démontré.

La relation (8) peut être utilisée de deux façons pour évaluer la proximité des solutions des problèmes différentiel et discret.

La première consiste à comparer directement (2) et (8). Pour fixer les idées, prenons le cas $a = b = 0$, $\delta_n \equiv 0$. Portons $x = nh$ et récrivons (2) sous la forme

$$y(nh) = \int_0^{nh} G(nh, s) f(s)\, ds + \int_{nh}^{X} G(nh, s) f(s)\, ds.$$

Si $p(x)$ et $f(x)$ sont deux fois continûment dérivables, on montre que les deux fonctions à intégrer possèdent également cette propriété. Remplaçons les deux intégrales par la formule des trapèzes de pas h. On trouve

$$y(nh) = h\left(\frac{G(nh, 0) f(0)}{2} + \sum_{k=1}^{n-1} G(nh, kh) f(kh) + \frac{G(nh, nh) f(nh)}{2}\right) +$$
$$+ h\left(\frac{G(nh, nh) f(nh)}{2} + \sum_{k=n+1}^{N-1} G(nh, kh) f(kh) + \right.$$
$$\left. + \frac{G(nh, Nh) f(Nh)}{2}\right) + O(h^2).$$

Comme $G(nh, 0) = G(nh, Nh) = 0$ par définition de la fonction de Green $G(x, s)$, on a

$$y(nh) = h \sum_{k=1}^{N-1} G(nh, kh) f_k + O(h^2). \tag{9}$$

On a effectué la bipartition de l'intégrale parce que l'erreur dans la formule des trapèzes est évaluée par la dérivée seconde de la fonction sous le signe $\int$, et la dérivée première de la fonction $G(nh, s)$ présente une discontinuité au point nh. Il découle des égalités (8), (9)

$$y_n - y(nh) = h \sum_{k=1}^{N-1} (G_n^k - G(nh, kh)) f_k + O(h^2).$$

Compte tenu de l'estimation (7) on obtient $y_n - y(nh) = O(h^2)$. On établit sans peine que cette estimation de l'erreur est uniforme en n pour $0 \leqslant x_n \leqslant X$, i.e.

$$\max_{0 \leqslant x_n \leqslant X} |y_n - y(nh)| = O(h^2).$$

La seconde, qui nous est plus familière, consiste à établir l'équation d'erreur et à évaluer l'erreur par la formule de Green. On sait (cf. § 1) que l'erreur R_n satisfait à

$$l(R_n) = \delta_n - r_n, \quad \text{où} \quad |r_n| \leqslant \frac{M_4 h^2}{12};$$

δ_n est l'erreur due à l'imprécision de la solution du système (1.3)-(1.4). Soit $R_0 = R_N = 0$. La fonction de Green du problème différentiel étant sous forme explicite, il en résulte qu'elle est uniformément bornée; compte tenu de (7) on trouve qu'il en est de même de G_n^k: $|G_n^k| \leqslant D$ pour $0 \leqslant kh$, $nh \leqslant X$. Explicitons R_n à l'aide de la formule de Green (9):

$$R_n = h \sum_{k=1}^{N-1} G_n^k (\delta_k - r_k). \tag{10}$$

Il en découle l'estimation

$$|R_n| \leqslant Dh \sum_{k=1}^{N-1} (|\delta_k| + |r_k|).$$

En réduisant la précision on obtient l'estimation plus suggestive:

$$|R_n| \leqslant DX \left(\frac{M_4}{12} h^2 + \max_{0<k<N} |\delta_k| \right).$$

Si $p(x) \geqslant 0$, celle-ci est du même ordre en h et en $\max |\delta_k|$ que l'estimation précédente. Les procédés décrits peuvent permettre d'évaluer l'erreur dans des hypothèses assez faibles sur la régularité des coefficients du problème (le premier procédé s'avère particulièrement utile en l'occurrence).

§ 3. Résolution d'un simple problème aux limites discrétisé

Dans la résolution numérique du problème de Cauchy, les valeurs de la solution sont définies aux nœuds successifs par des formules de récurrence; un problème aux limites discrétisé tel que (1.3)-(1.4) ne présente pas cette possibilité puisque les valeurs de la solution dépendent des conditions aux limites aux deux extrémités du segment d'intégration.

Le problème aux limites

$$y'' - p(x)\, y = f(x), \quad y(0) = a, \quad y(X) = b$$

peut être résolu comme suit. Prenons une solution particulière $y_0(x)$ de l'équation avec second membre

$$y_0'' - p(x)\, y_0 = f(x)$$

et deux solutions linéairement indépendantes de l'équation homogène $y_i'' - p(x)\, y_i = 0$, $i = 1, 2$. La solution générale de l'équa-

tion non homogène s'écrit alors

$$y_0(x) + C_1 y_1(x) + C_2 y_2(x),$$

avec C_1 et C_2 constants définis par les conditions aux limites. Approchons les fonctions $y_i(x)$, $i = 0, 1, 2$, par une méthode de résolution numérique du problème de Cauchy; définissons C_i et trouvons ainsi la solution cherchée.

Il est possible de procéder plus économiquement. A cette fin on trouve une solution particulière de l'équation non homogène $y_0'' - p(x) y_0 = f(x)$, qui vérifie la condition $y_0(0) = a$, et une solution particulière $y_1(x)$ de l'équation homogène telle que $y_1(0) = 0$. La solution générale de l'équation non homogène, qui satisfait à $y(0) = a$, s'écrit $y_0(x) + Cy_1(x)$; la valeur de C est définie moyennant la condition $y_0(X) + Cy_1(X) = b$.

La méthode de résolution qui s'inspire de ce schéma porte le nom de *méthode du tir*. Voici l'idée de son analogue discret. On se donne $y_0^0 = a$, $y_0^1 = 0$, $y_1^1 \neq 0$ et y_1^0 arbitraires, et on détermine de proche en proche $y_2^0, \ldots, y_N^0$, $y_2^1, \ldots, y_N^1$ à l'aide de

$$\frac{y_{n+1}^0 - 2y_n^0 + y_{n-1}^0}{h^2} - p_n y_n^0 = f_n, \tag{1}$$

$$\frac{y_{n+1}^1 - 2y_n^1 + y_{n-1}^1}{h^2} - p_n y_n^1 = 0. \tag{2}$$

On trouve ensuite C à partir de $y_N^0 + Cy_N^1 = b$ et on pose $y_n = y_n^0 + Cy_n^1$; la fonction y_n est la solution cherchée.

S'il est impossible de loger toutes ces quantités en mémoire rapide de la machine, au lieu de mettre en réserve y_n^i, $i = 0, 1$, on recherche, une fois C calculé, $y_1 = y_1^0 + Cy_1^1$, puis on détermine successivement $y_2, \ldots, y_{N-1}$ moyennant $l(y_n) = f_n$. L'algorithme décrit s'applique formellement avec y_1^0 et y_1^1 quelconques; cependant, la valeur $y_1^0 = a + O(h)$ est la plus raisonnable si l'on veut diminuer l'erreur de calcul.

Justifions notre dernière assertion. Dans les calculs effectifs, les valeurs $\overline{y}_n^0$ sont reliées par

$$\overline{y}_{n+1}^0 = 2\overline{y}_n^0 - \overline{y}_{n-1}^0 + h^2 (p_n \overline{y}_n^0 - f_n) + \delta_n;$$

le résultat est arrondi en calcul automatique avec une erreur relative de l'ordre de 2^{-t}, et δ_n peut être de l'ordre de $2^{-t}\overline{y_{n+1}^0}$. Introduisons les fonctions auxiliaires: y_n^2, solution de (1) avec les conditions initiales $y_0^2 = y_1^2 = a$ et y_n^3, solution de (2) avec pour conditions initiales $y_0^3 = 0$, $y_1^3 = h$. Soit $y^2(x)$ une solution de $y'' - p(x) y = f(x)$ vérifiant les conditions initiales $y(0) = a$, $y'(0) = 0$. Nous avons montré au § 2 que

$$\max_{0 \leqslant x_n \leqslant X} |y_n^3 - W^1(nh)| \leqslant Mh^2; \tag{3}$$

en évaluant par le même procédé la proximité entre les solutions y_n^2 et $y^2(nh)$ des problèmes discret et différentiel respectivement, on aboutit, à l'aide du

théorème du § 14, chap. VIII, à

$$\max_{0 \leqslant x_n \leqslant X} | y_n^2 - y^2(nh) | = O(h). \tag{4}$$

La fonction $y_n^2 + y_n^3 (y_1^0 - a) h^{-1}$ coïncide avec y_n^0 pour $n = 0, 1$ et vérifie la même équation aux différences finies; aussi $y_n^0 = y_n^2 + y_n^3 (y_1^0 - a) h^{-1}$. Compte tenu des estimations (3)-(4) il vient:

$$y_n^0 = (y^2(nh) + O(h)) + (W^1(nh) + O(h)) (y_1^0 - a) h^{-1}.$$

Il s'ensuit que, lorsque $| (y_1^0 - a) h^{-1} | \gg 1$, les quantités y_n^0 sont de l'ordre de $(y_1^0 - a) h^{-1}$. Puisque les valeurs y_n^0 et $\overline{y}_n^0$ diffèrent relativement peu, les $\overline{y}_1^0$ sont également du même ordre de grandeur. Conclusion: on choisira de préférence y_1^0 tel que $(y_1^0 - a) h^{-1}$ soit petit.

De nombreux problèmes aux limites se caractérisent par une croissance notable (avec la croissance ou la diminution de x) de certaines des solutions des équations différentielles homogènes correspondantes.

Considérons l'équation modèle

$$y'' - py = 0, \; p = \text{const} > 0. \tag{5}$$

Les solutions $\exp(\pm\sqrt{p}x)$ de l'équation homogène associée croissent (resp. décroissent) brusquement sur le tronçon considéré si $\sqrt{p}X$ est suffisamment élevé. $\sqrt{p}X$ et h sont donc les paramètres qui déterminent l'influence de l'erreur de calcul dans les différentes méthodes de résolution du problème discret (1.3)-(1.4). Examinons le modèle d'accumulation de l'erreur de calcul à l'une des étapes de résolution du problème discret. Supposons qu'on a trouvé une valeur y_1 exempte de toute erreur et qu'ensuite on calcule à l'aide de la formule de récurrence

$$y_{n+1} = 2y_n - y_{n-1} + ph^2 y_n.$$

Supposons qu'il existe la relation

$$y_{n+1}^* = 2y_n^* - y_{n-1}^* + ph^2 y_n^* + \delta$$

(δ étant de l'ordre de 2^{-t}) entre les y_n^* obtenus réellement. En en retranchant l'égalité précédente, on aboutit à une équation donnant l'erreur $\Delta_n = y_n^* - y_n$:

$$\Delta_{n+1} = 2\Delta_n - \Delta_{n-1} + ph^2\Delta_n + \delta, \tag{6}$$

avec, par hypothèse, $\Delta_0 = \Delta_1 = 0$. L'équation caractéristique

$$\mu^2 - (2 + ph^2)\mu + 1 = 0$$

de cette équation aux différences possède les racines

$$\mu_{1,2} = 1 + \frac{ph^2}{2} \pm \sqrt{ph^2 + \frac{(ph^2)^2}{4}}.$$

Cherchons une solution particulière de l'équation non homogène sous la forme $\Delta_n = c$. Son report dans (6) donne $c = -\frac{\delta}{ph^2}$. Ainsi, la solution générale de l'équation non homogène s'écrit

$$\Delta_n = C_1\mu_1^n + C_2\mu_2^n - \frac{\delta}{ph^2}; \tag{7}$$

nous définissons C_1, C_2 à partir des conditions $\Delta_0 = \Delta_1 = 0$. Donc, finalement,

$$\Delta_n = \frac{\delta}{ph^2}\left(\frac{1-\mu_2}{\mu_1-\mu_2}\mu_1^n + \frac{1-\mu_1}{\mu_2-\mu_1}\mu_2^n - 1\right).$$

Admettons que $\sigma = \sqrt{ph^2}$ est petit. Alors

$$\mu_{1,2} = 1 \pm \sigma + O(\sigma^2),$$

et, par conséquent,

$$\frac{1-\mu_2}{\mu_1-\mu_2} = \frac{\sigma + O(\sigma^2)}{2\sigma + O(\sigma^2)} = \frac{1}{2} + O(\sigma);$$

de même

$$\frac{1-\mu_1}{\mu_2-\mu_1} = \frac{1}{2} + O(\sigma).$$

Comme d'ailleurs $\mu_{1,2} = \exp(\pm\sigma + O(\sigma^2))$, l'égalité (7) se met sous la forme

$$\Delta_n = \frac{\delta}{ph^2}\left(\left(\frac{1}{2} + O(\sigma)\right)\exp(\sigma n(1 + O(\sigma))) + \right.$$
$$\left. + \left(\frac{1}{2} + O(\sigma)\right)\exp(-\sigma n(1 + O(\sigma))) - 1\right).$$

Lorsque $nh = x$, on a $\sigma n = \sqrt{p}\,x$ et

$$\Delta_n = \frac{\delta}{ph^2}\left(\left(\frac{1}{2} + O(\sigma)\right)\exp(\sqrt{p}\,x(1 + O(\sigma))) + \right.$$
$$\left. + \left(\frac{1}{2} + O(\sigma)\right)\exp(-\sqrt{p}\,x(1 + O(\sigma))) - 1\right). \tag{8}$$

Si $\sqrt{p}x$ est grand, pour x voisin de X, l'influence de l'erreur de calcul est catastrophique conformément à (8). Fait remarquable, dans de nombreux cas, la cause en est non pas le facteur h^{-2}, conséquence de la petitesse du pas, mais $\exp(\sqrt{p}x(1 + O(\sigma)))$. Par exemple, dans le problème aux limites pour $y'' - y = f$ sur $[0, 30]$ on a $\exp(\sqrt{1}\cdot 30) = 10^{12,\cdots}$, tandis qu'on recourt rarement à un pas d'intégration $h < 10^{-3}$.

Lorsque $p(x) \geqslant 0$, on fera appel, pour le système (1.3)-(1.4), à la *méthode du balayage*. Représentons la condition aux limites $y_0 = a$ par $y_0 = C_0 y_1 + \varphi_0$, où $C_0 = 0$, $\varphi_0 = a$. Portant $y_0 =$

$= C_0 y_1 + \varphi_0$ dans la première équation du système (1.3)-(1.4)

$$\frac{y_0 - 2y_1 + y_2}{h^2} - p_1 y_1 = f_1,$$

on trouve une relation entre y_1 et y_2. Sa résolution par rapport à y_1 donne

$$y_1 = C_1 y_2 + \varphi_1, \tag{9}$$

avec

$$C_1 = \frac{1}{2 + p_1 h^2}, \quad \varphi_1 = C_1 (a - f_1 h^2);$$

en substituant l'expression de y_1 en fonction de y_2 dans la deuxième équation du système on trouve une relation entre y_2 et y_3, et ainsi de suite. Soit, enfin,

$$y_n = C_n y_{n+1} + \varphi_n; \tag{10}$$

portons cette expression de y_n dans la $(n+1)$-ième équation du système (1.3)-(1.4):

$$\frac{y_n - 2y_{n+1} + y_{n+2}}{h^2} - p_{n+1} y_{n+1} = f_{n+1}.$$

Après avoir résolu l'équation ainsi obtenue

$$\frac{(C_n y_{n+1} + \varphi_n) - 2y_{n+1} + y_{n+2}}{h^2} - p_{n+1} y_{n+1} = f_{n+1},$$

on a

$$y_{n+1} = C_{n+1} y_{n+2} + \varphi_{n+1},$$

où

$$C_{n+1} = \frac{1}{2 + p_{n+1} h^2 - C_n}, \quad \varphi_{n+1} = C_{n+1} (\varphi_n - f_{n+1} h^2). \tag{11}$$

On voit que les coefficients des équations (10) reliant deux valeurs consécutives y_n et y_{n+1} s'obtiennent à partir des relations récurrentes (11) dans les conditions initiales $C_0 = 0$, $\varphi_0 = a$. Vu que y_N est connu, on détermine de proche en proche, avec les coefficients C_n, φ_n trouvés, $y_{N-1}, \ldots, y_1$ à l'aide de (10). Le calcul des coefficients C_n, φ_n s'appelle *balayage en sens direct* et celui des inconnues y_n *balayage en sens inverse* ou *par remontée*. La marche des opérations est schématisée comme suit:

balayage en sens direct

$$\begin{array}{ccccccccc}
 & & p_1 & & p_2 & & & & p_{N-1} \\
 & & \downarrow & & \downarrow & & & & \downarrow \\
C_0 = 0 & \to & C_1 & \to & C_2 & \to & \ldots & \to & C_{N-1} \\
 & & \downarrow & & \downarrow & & & & \downarrow \\
\varphi_0 = a & \to & \varphi_1 & \to & \varphi_2 & \to & \ldots & \to & \varphi_{N-1} \\
 & & \uparrow & & \uparrow & & & & \uparrow \\
 & & f_1 & & f_2 & & & & f_{N-1}
\end{array},$$

balayage par remontée :

$$\begin{array}{ccccccc} C_1 & & C_2 & & & & C_{N-1} \\ \downarrow & & \downarrow & & & & \downarrow \\ y_1 & \leftarrow & y_2 & \leftarrow & \ldots & \leftarrow & y_{N-1} \leftarrow y_N. \\ \uparrow & & \uparrow & & & & \uparrow \\ \varphi_1 & & \varphi_2 & & & & \varphi_{N-1} \end{array}$$

L'appellation de la méthode du balayage est parfois justifiée par les raisonnements suivants. L'équation

$$y_n = C_n y_{n+1} + \varphi_n$$

est la conséquence de la condition à la limite $x = 0$ et des équations du système (1.3) correspondant aux points $x_j \leqslant x_n$. Cette égalité est donc vraie pour toute solution du système d'équations $l(y_j) = f_j$, $j = 1, \ldots, n$, qui vérifie la condition à l'extrémité gauche; la condition à la limite $x = 0$ est pour ainsi dire « balayée » jusqu'au point courant $x = x_n$.

Les méthodes du tir et du balayage sont des cas particuliers de l'algorithme d'élimination de Gauss pour le système (1.3)-(1.4). En ramenant n équations de ce système à la forme (10) par le balayage en sens direct on obtient le système

$$\begin{array}{ll} y_1 - C_1 y_2 & = \varphi_1, \\ \vdots & \vdots \\ y_n - C_n y_{n+1} & = \varphi_n, \\ y_n - (2 + p_{n+1}h^2)\, y_{n+1} + y_{n+2} & = f_{n+1}h^2 \\ \vdots & \vdots \end{array}$$

résultat auquel on aboutirait par élimination de n inconnues dans la marche directe de la méthode de Gauss. Le balayage en sens direct réduit la matrice du système à une forme triangulaire. Le balayage par remontée coïncide avec la méthode de Gauss de remontée.

PROBLÈME. Ecrire les formules de calcul de la méthode de la racine carrée pour le système considéré. Comparer le coût de ce procédé avec le travail qu'exige la méthode du balayage.

Examinons une autre variante de la méthode de Gauss pour le système (1.3)-(1.4). Exprimons, à l'aide de la première équation, y_2 en fonction de y_1 et substituons dans les autres. La deuxième équation ne renferme maintenant que les inconnues y_1 et y_3; exprimons y_3 moyennant y_1 et substituons dans les équations restantes, et ainsi de suite. Supposons que nous avons déjà exprimé $y_2, \ldots, y_n$ à l'aide de y_1 et trouvé les relations $y_j = \alpha_j y_1 + \beta_j$ pour $2 \leqslant j \leqslant n$; par souci d'uniformité ajoutons-y les égalités

$$y_0 = \alpha_0 y_1 + \beta_0, \text{ où } \alpha_0 = 0,\ \beta_0 = a,$$
$$y_1 = \alpha_1 y_1 + \beta_1, \text{ où } \alpha_1 = 1,\ \beta_1 = 0.$$

Le report des expressions de y_{n-1} et de y_n dans l'équation

$$y_{n-1} - (2 + p_n h^2)\, y_n + y_{n+1} = f_n h^2$$

fournit

$$(\alpha_{n-1} y_1 + \beta_{n-1}) - (2 + p_n h^2)\,(\alpha_n y_1 + \beta_n) + y_{n+1} = f_n h^2,$$

ou encore

$$y_{n+1} = \alpha_{n+1} y_1 + \beta_{n+1},$$

où

$$\begin{aligned} \alpha_{n+1} &= (2 + p_n h^2)\,\alpha_n - \alpha_{n-1}, \\ \beta_{n+1} &= (2 + p_n h^2)\,\beta_n - \beta_{n-1} + f_n h^2. \end{aligned} \tag{12}$$

Ainsi, les coefficients α_n, β_n se calculent de proche en proche par les formules de récurrence (12). Après avoir obtenu $y_N = \alpha_N y_1 + \beta_N$, déterminons la valeur y_1, puis tous les y_j par

$$y_j = \alpha_j y_1 + \beta_j. \tag{13}$$

Vu (12), les valeurs α_j satisfont à l'équation aux différences finies homogène (2) et aux conditions initiales $\alpha_0 = 0$, $\alpha_1 = 1$, et les valeurs β_j à l'équation aux différences finies non homogène (1) et aux conditions initiales $\beta_0 = a$, $\beta_1 = 0$. Dans la méthode du tir, pour un certain choix des conditions initiales, α_n se confond avec y_n^1 et β_n avec y_n^0. La fonction $z_n = \alpha_n C + \beta_n$ vérifie l'équation aux différences finies non homogène et la condition à l'extrémité gauche pour n'importe quel C, de plus $z_1 = \alpha_1 C + \beta_1 = C$. Par conséquent, en résolvant $y_N = \alpha_N C + \beta_N$ par rapport à C, nous tombons sur la valeur $C = y_1$ recherchée par la méthode du tir. Les calculs selon (13) correspondent à la recherche des y_n par la formule $y_n = y_n^1 C + y_n^0$. La méthode construite coïncide avec l'une des mises en œuvre de la méthode du tir et avec l'algorithme de Gauss si l'ordre d'élimination des inconnues est $y_0, y_2, \ldots, y_N, y_1$.

La résolution d'un système d'équations ayant une $N \times N$-matrice A pleine par l'algorithme de Gauss classique exige $\asymp N^3$ opérations arithmétiques.

Les matrices $A = [a_{ij}]$, $a_{ij} = 0$ pour $|\, i - j \,| > m$, sont dites $(2m + 1)$-*diagonales*; dans le cas de matrices tridiagonales, l'élimination de Gauss qui coïncide avec la méthode du balayage demande, nous l'avons vu, un nombre sensiblement moindre d'opérations. De même, ce nombre est notablement plus faible pour les matrices $(2m + 1)$-diagonales si $m \ll N$.

Voici un problème type qui se réduit à un système d'équations de matrice proche d'une forme $(2n + 1)$-diagonale.

Le lissage des fonctions par la régularisation (§ 4, chap. V) a été à l'origine du problème suivant. Soit une fonction périodique f_q de l'argument q entier; la période est égale à $N + 1$ (N au § 4, ch. V). Trouver une fonction périodique u_q de même période satisfaisant à

$$\frac{\delta^{2n} u_q}{h^{2n}} + (-\lambda^2)^n u_q = f_q \quad \text{quels que soient } q.$$

Ecrivons ces relations pour $q = 0, \ldots, N$. Vu la périodicité, on remplace les valeurs de u_q avec $q < 0$ et $q > N$ figurant dans ce système par u_{q+N+1} et u_{q-N-1} respectivement. D'où un système de $N + 1$ équations par rapport aux inconnues $u_0, \ldots, u_N$: $Au = f$. Les éléments de la matrice $A = [a_{ij}]$ du système se définissent par $a_{ij} = a(|i - j|)$, avec

$$a(k) = \begin{cases} a(0) = (-1)^n (C_{2n}^n h^{-2n} + \lambda^{2n}), \\ (-1)^{n-k} C_{2n}^{n-k} h^{-2n} & \text{pour } 0 < k \leqslant n, \\ 0 & \text{pour } n < k \leqslant N-n, \\ (-1)^{k+n-N-1} C_{2n}^{k+n-N-1} h^{-2n} & \text{pour } N-n < k \leqslant N \end{cases}$$

$$A = \begin{pmatrix} a(0) & a(1) & \ldots & a(n) & 0 & \ldots & 0 & a(N-n+1) & \ldots & a(N) \\ a(1) & & & & & & & & & \vdots \\ \vdots & & & & & & & & & \\ a(n) & & & & & & & & & \vdots \\ 0 & & & & & & & & & \\ \vdots & & & & & & & & & \\ 0 & & & & & & & & & \vdots \\ a(N-n+1) & & & & & & & & & \\ \vdots & & & & & & & & & \\ a(N) & \ldots & & \ldots & & \ldots & & & & \end{pmatrix}$$

La matrice A est symétrique et définie positive.

Ce système peut se résoudre par la méthode de la racine carrée. L'autre procédé consiste à éliminer successivement les inconnues écrites dans l'ordre naturel $u_1, u_2, \ldots$ Dans les deux cas, la transformation du système s'exprime par des formules de récurrence analogues à celles de la méthode du balayage, et le nombre d'opérations arithmétiques nécessaires reste $O(n^2N)$. Puisqu'on recourt à des opérations différentes, le choix de la méthode tient compte du coût relatif en opérations machine. L'erreur de calcul atteint dans les deux cas des valeurs de l'ordre de $2^{-t}h^{-2n}$.

PROBLÈMES. 1. Ecrire les formules de calcul des deux méthodes pour $n = 1$.

2. Quel est le nombre d'opérations arithmétiques exigées par les deux méthodes avec n quelconque?

Dans certaines situations la matrice A du système est bloc-$(2n + 1)$-diagonale, i.e. elle est partitionnée en blocs A_{ij} tels que $A_{ij} \equiv 0$ lorsque $|i - j| > n$. On utilise alors la méthode d'élimination par groupes correspondant aux blocs, et les formules de calcul sont les mêmes que pour les systèmes de matrice $(2n + 1)$-diagonale (la seule différence est qu'il faut remplacer les opérations scalaires par les matricielles).

Prenant comme modèle le cas $p(x) \equiv p = \text{const}$, examinons le comportement des coefficients de balayage C_n pour p de signes différents. Les relations

$$C_{n+1} = \frac{1}{2 + ph^2 - C_n},$$

qui sont vérifiées par ces coefficients, coïncident avec les formules d'itération pour

$$C = g(C) = \frac{1}{2 + ph^2 - C}$$

avec l'approximation au départ $C_0 = 0$. L'équation $C = g(C)$ est équivalente à l'équation du second degré

$$C^2 - (2 + ph^2)\, C + 1 = 0; \tag{14}$$

ses racines sont

$$C_{(1)},\ C_{(2)} = 1 + \frac{ph^2}{2} \pm \sqrt{ph^2 + \frac{p^2h^4}{4}}.$$

Pour $ph^2 > 0$, on a évidemment $C_{(1)} > 1$, donc $0 < C_{(2)} = C_{(1)}^{-1} < 1$. Pour $p < 0$ et h petit, l'expression sous le radical est négative

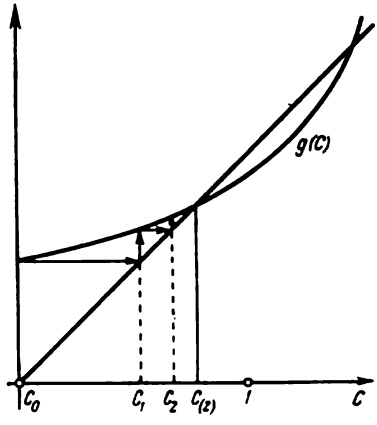


Fig. 9.3.1

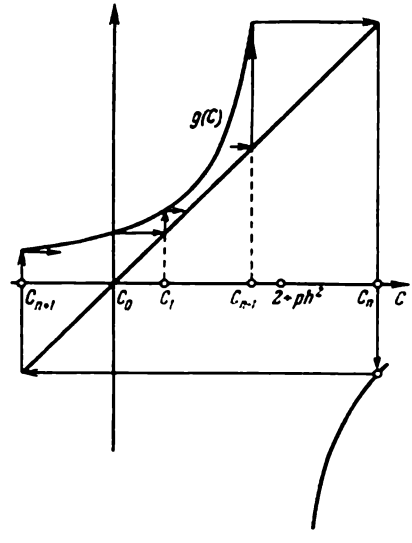


Fig. 9.3.2

et l'équation (14) n'admet pas de racines réelles. Les fig. 9.3.1-9.3.2 représentent la position relative des courbes $y = C$ et $y = g(C)$ et des points (C_n, C_n), (C_n, C_{n+1}). On voit que lorsque $p \geqslant 0$, les valeurs de C_n sont dans les limites $0 < C_n \leqslant 1$ (fig. 9.3.1). Si $p < 0$, il n'est pas exclu (fig. 9.3.2) que, pour n suffisamment important,

une valeur C_{n-1} soit voisine de $2 + ph^2$; la valeur suivante C_n est alors très élevée. Cette circonstance nous oblige à étudier plus en détail l'influence de l'erreur de calcul, au moins pour $p < 0$.

§ 4. Fermetures des algorithmes de calcul

La résolution d'un problème modèle fournit de nombreux renseignements utiles pour l'analyse préliminaire d'un algorithme. Un autre moyen efficace d'analyse est l'étude des fermetures des algorithmes de calcul.

Afin de ne pas alourdir l'exposé nous nous bornerons à l'essence du problème sans justifier suffisamment certaines constructions du présent paragraphe.

Soit un problème

$$Lu = f, \tag{1}$$

et soit

$$L^h u^h = f^h \tag{2}$$

une suite de problèmes dépendant du paramètre h (cela peut être le pas du réseau) dont les solutions u^h convergent pour $h \to 0$ vers la solution du problème initial. Supposons que l'algorithme de résolution de (2) consiste à obtenir de proche en proche des relations

$$L_m^h u_m^h = f_m^h, \quad m = 1, \ldots, M; \tag{3}$$

$L_M^h = E$ est l'opérateur unité, $u_M^h = f_M^h = (L^h)^{-1} f^h = u^h$, c'est-à-dire qu'on obtient au M-ième pas la solution exacte u^h. Introduisons un paramètre $z(m, h)$, fonction monotone de m et tel que pour $h \to 0$ et z fixe la relation (3) devienne à la limite

$$L_z y = f_z, \quad 0 \leqslant z \leqslant z_0, \tag{4}$$

où

$$z_0 = \lim_{h \to 0} z(M, h).$$

La relation (4) s'appelle *fermeture de l'algorithme de calcul* (3).

Pour l'instant, nous ne recommandons pas au lecteur d'approfondir cette notion peu rigoureuse. Celle-ci sera plus claire quand nous l'utiliserons pour le problème aux limites.

La définition de l'algorithme de calcul n'exclut pas le cas $M = \infty$. Dans ce cas les égalités

$$L_M^h = E, \quad f_M^h = (L^h)^{-1} f^h$$

s'interprètent comme suit :

$$L_m^h \to E, \quad f_m^h \to (L^h)^{-1} f^h \quad \text{pour} \quad m \to \infty.$$

Le cas $M = \infty$ correspond aux méthodes itératives de résolution du problème (1). Si les opérateurs L_z sont bornés uniformément en z, on dit que l'algorithme (3) possède une *fermeture régulière*; dans le cas contraire, cette fermeture est *irrégulière*.

Si la fermeture est régulière, on a des raisons de penser que l'algorithme sera insensible à diverses perturbations, en particulier à l'erreur de calcul. L'étude de la fermeture constitue donc un moyen commode de se faire une idée de la qualité d'un nouveau procédé dès le départ. Cette évaluation préliminaire n'est pas forcément définitive. Il se peut que les opérateurs L_z soient uniformément bornés par une constante très grande, ce qui en fait revient à dire qu'ils ne sont pas bornés. D'autre part, il est possible que les opérateurs L_z ne sont pas bornés à cause d'une normalisation défectueuse des équations (1.3) ou encore d'un choix maladroit des normes dans les espaces de fonctions u. Le fait, pour L_z, de ne pas être uniformément bornés signifie que $q(h) = \sup_m \| L_m^h \|$ croît indéfiniment lorsque h tend vers zéro. Il semble que la présence de grands coefficients dans $L_m^h u_m^h = f_m^h$ devrait être équivalente à des perturbations notables des coefficients du problème initial et, partant, à une influence sérieuse de l'erreur de calcul. Il se peut cependant que $q(h)$ croisse très lentement et que l'augmentation correspondante de l'influence de l'erreur de calcul soit acceptable lorsque le pas tend vers zéro. Malgré ces considérations l'étude des fermetures des algorithmes de calcul s'avère fort utile.

Avec le balayage en sens direct on a

$$y_n - C_n y_{n+1} = \varphi_n. \tag{5}$$

Voyons ce que deviennent ces relations lorsque $h \to 0$. Prenons à cette fin le cas le plus simple de $p \equiv 0$. Alors $C_0 = 0$, $C_1 = 1/2$, $C_2 = 2/3$, ... et (on l'établit facilement en raisonnant par récurrence) $C_n = 1 - 1/(n+1)$. Fixons un point quelconque, disons $x = nh$. Le premier membre de (5) qui lui correspond est

$$y_n - \left(1 - \frac{1}{\frac{x}{h}+1}\right) y_{n+1}.$$

En remplaçant y_n par les valeurs de la fonction régulière $y(nh)$ la limite du premier membre est égale à zéro, et l'étude d'une telle fermeture de l'algorithme n'a pas de sens. Pour qu'il n'en soit pas ainsi, on choisit un $\lambda(h)$ tel que la limite

$$\lim_{h \to 0} \lambda(h) \left(y(x) - \left(1 - \frac{1}{\frac{x}{h}+1}\right) y(x+h) \right) \tag{6}$$

soit finie et non nulle. Soit $\lambda(h) = h^{-1}$; alors

$$\frac{1}{h}\left(y(x)-\left(1-\frac{1}{\frac{x}{h}+1}\right)y(x+h)\right)=$$

$$=\frac{1}{h}\left(y(x)-\left(1-\frac{h}{x}+O(h^2)\right)(y(x)+y'(x)\,h+O(h^2))\right)=$$

$$=\frac{y(x)}{x}-y'(x)+O(h).$$

Dans une telle normalisation le premier membre de (6) tend vers une limite différente de zéro. Essayons de nous servir de cette normalisation dans le cas général. Puisque dans l'exemple considéré il existait

$$\lim_{h\to 0}\frac{C_n-1}{h} \quad \text{lorsque } nh=x,$$

introduisons au lieu de C_n une nouvelle variable $\alpha_n=\frac{1-C_n}{h}$ et rappelons les relations de la méthode du balayage

$$C_{n+1}=\frac{1}{2+p_{n+1}h^2-C_n}, \quad \varphi_{n+1}=C_{n+1}(\varphi_n-f_{n+1}h^2). \tag{7}$$

Multiplions (5) par $-h^{-1}$ et posons $C_n=1-\alpha_n h$ et $\varphi_n=-\beta_n h$. La relation (5) se récrit

$$\frac{y_{n+1}-y_n}{h}-\alpha_n y_{n+1}=\beta_n. \tag{8}$$

Portons C_n et φ_n exprimés par α_n et β_n dans (7); il vient

$$1-\alpha_{n+1}h=\frac{1}{1+p_{n+1}h^2+\alpha_n h},$$

$$\beta_{n+1}h=(1-\alpha_{n+1}h)(\beta_n h+f_{n+1}h^2),$$

d'où

$$\alpha_{n+1}=\frac{\alpha_n+p_{n+1}h}{1+\alpha_n h+p_{n+1}h^2}, \tag{9}$$

$$\beta_{n+1}=\beta_n+(f_{n+1}-\alpha_{n+1}\beta_n)\,h-\alpha_{n+1}f_{n+1}h^2. \tag{10}$$

L'égalité (9) se transforme en

$$\frac{\alpha_{n+1}-\alpha_n}{h}=p_{n+1}-\alpha_n^2-h\,\frac{2p_{n+1}\alpha_n-\alpha_n^3+(p_{n+1}^2-p_{n+1}\alpha_n^2)\,h}{1+\alpha_n h+p_{n+1}h^2}, \tag{11}$$

et (10) en

$$\frac{\beta_{n+1}-\beta_n}{h}=f_{n+1}-\alpha_{n+1}\beta_n-f_{n+1}\alpha_{n+1}h. \tag{12}$$

Sous l'hypothèse des coefficients α_n et β_n tendant, pour $nh = x$ fixe et $h \to 0$, vers des limites $\alpha(x)$ et $\beta(x)$, la substitution des valeurs

$y(nh)$ à y_n dans (8) donne à la limite l'équation différentielle

$$y' - \alpha(x)\, y = \beta(x),$$

qui est la fermeture de l'algorithme de calcul. Les relations (11), (12) montrent que α_n, β_n vérifient des équations qui font penser à la méthode d'intégration numérique des équations différentielles d'Euler:

$$\alpha' = p(x) - \alpha^2, \tag{13}$$

$$\beta' = f(x) - \alpha\beta. \tag{14}$$

En prouvant le fait que α_n, β_n tendent, pour $nh = x$ fixe et $h \to 0$, vers les solutions de ces équations on se heurte à la difficulté suivante. Par définition,

$$\alpha_0 = \frac{1 - C_0}{h} = \frac{1}{h}, \quad \beta_0 = -\frac{\varphi_0}{h} = -\frac{a}{h}. \tag{15}$$

Ainsi, dans le voisinage du point $x = 0$ les conditions initiales de l'intégration numérique par les formules (11), (12) sont assez spéciales si bien que dans ce cas on ne peut appliquer les méthodes que nous avons développées pour évaluer la proximité des solutions des problèmes discret et différentiel. La forme de (15) autorise à supposer que $\alpha(x)$ est une solution de (13), qui se comporte à proximité de zéro comme x^{-1}, et $\beta(x)$ une solution de (14), qui se comporte dans le même domaine comme $-ax^{-1}$. Justifions nos affirmations.

Les coefficients C_n et donc α_n s'obtiennent des relations de récurrence (7) indépendantes du second membre de l'équation différentielle et de la condition aux limites au point X; il en est de même de la condition initiale $C_0 = 0$ de la récurrence. Considérons le problème aux limites

$$y'' - p(x)\, y = 0, \quad y(0) = 0, \quad y(X) = 1$$

et le problème discret correspondant

$$l(y_j) = 0, \quad y_0 = 0, \quad y_N = 1.$$

Leurs solutions s'écrivent

$$y(x) = \frac{W^1(x)}{W^1(X)}, \quad y_n = \frac{W_n^1}{W_N^1}$$

(cf. § 2). Comme $\varphi_0 = 0$ et tous les $f_n = 0$ pour notre problème aux limites, on a $\beta_n \equiv 0$ et la forme suivante

$$\frac{y_{n+1} - y_n}{h} - \alpha_n y_{n+1} = 0$$

de la relation (8). En y portant $y_n = W_n^1/W_N^1$, il vient

$$\alpha_n = \left(\frac{W_{n+1}^1 - W_n^1}{h}\right)(W_{n+1}^1)^{-1}. \tag{16}$$

On a l'estimation (§ 2)

$$|\, W_n^1 - W^1(nh)\,| \leqslant Mh^2 \text{ pour } 0 \leqslant nh \leqslant X, \tag{17}$$

$(W^1)''_{xx}$ est uniformément borné sur $[0, X]$; c'est pourquoi

$$\frac{W^1_{n+1}-W^1_n}{h}=\frac{W^1((n+1)h)-W^1(nh)+O(h^2)}{h}=(W^1)'((n+1)h)+O(h) \qquad (18)$$

uniformément sur $[0, X]$. La substitution de ces expressions dans (16) fournit

$$\alpha_n=\frac{(W^1)'((n+1)h)+O(h)}{W^1((n+1)h)+O(h)}.$$

Ainsi, lorsque $h\to 0$, les valeurs α_n se rapprochent de celles de la fonction

$$\alpha(x)=\frac{(W^1)'(x)}{W^1(x)}.$$

Celle-ci satisfait à l'équation différentielle (13) vu que

$$\left(\frac{(W^1)'}{W^1}\right)'=\frac{(W^1)''}{W^1}-\left(\frac{(W^1)'}{W^1}\right)^2=p(x)-\alpha^2.$$

D'après la définition de $W^1(x)$, pour x petits $W^1(x)\sim x$ et $(W^1)'(x)\sim 1$ et, par conséquent, $\alpha(x)=\frac{(W^1)'(x)}{W(x)}\sim x^{-1}$, ce qui correspond à l'hypothèse. Notons que la fonction $\alpha(x)$ devient infinie aux points en lesquels $W^1(x)=0$, en particulier lorsque $x=0$. Dans le voisinage de chaque y pareil

$$(W^1)'(x)\sim(W^1)'(y)\neq 0, \quad W^1(x)\sim((W^1)'(y))(x-y)$$

et donc

$$\alpha(x)\sim(x-y)^{-1}. \qquad (19)$$

Les points de ce voisinage vérifient

$$\frac{W^1_{n+1}-W^1_n}{h}=(W^1)'((n+1)h)(1+O(h)),$$

$$W^1_{n+1}=W^1((n+1)h)+O(h^2)=W^1((n+1)h)\left(1+O\left(\frac{h^2}{W^1((n+1)h)}\right)\right)=$$

$$=W^1((n+1)h)\left(1+O\left(\frac{h^2}{(n+1)h-y}\right)\right),$$

ainsi

$$\alpha_n=\alpha((n+1)h)(1+O(h))\left(1+O\left(\frac{h^2}{(n+1)h-y}\right)\right).$$

L'erreur relative dans $\alpha_n\approx\alpha((n+1)h)$ est en $O(h)$ jusqu'au voisinage d'ordre h des points en lesquels $\alpha(x)=\infty$.

Selon (8), on a pour une équation non homogène

$$\beta_n=\frac{y_{n+1}-y_n}{h}-\alpha_n y_{n+1}.$$

Par suite de (18) et de l'estimation de la convergence de la solution du problème discret vers celle du problème différentiel (§ 2) on a

$$\beta_n=\beta((n+1)h)+O(h)$$

en tous les points $x=(n+1)h$, où $\alpha(x)\neq 0$; ici

$$\beta(x)=y'(x)-\alpha(x)y(x).$$

La fonction discrète β_n se rapproche donc d'une fonction $\beta(x)$ de l'argument continu. Evidemment,

$$\beta' = y'' - \alpha' y - \alpha y' = (py + f) - (p - \alpha^2)\, y - \alpha y' = \\ = f - \alpha (y' - \alpha y) = f - \alpha\beta;$$

au voisinage de 0, on a, vu (19),

$$\beta(x) \sim -y(0)\, x^{-1} \sim -ax^{-1}.$$

Ecrivons les relations (3) et les fermetures correspondantes. Posons $M = 2N - 1$ et acceptons pour (3), lorsque $m = 1, \ldots, N$, l'ensemble des relations

$$\frac{y_j - C_j y_{j+1}}{h} = \frac{\varphi_j}{h}, \quad 0 \leqslant j < m,$$

$$\frac{y_{j-1} - 2y_j + y_{j+1}}{h^2} - p_j y_j = f_j, \quad m \leqslant j < N,$$

$$y_N = b,$$

et lorsque $N < m \leqslant M$,

$$\frac{y_j - C_j y_{j+1}}{h} = \frac{\varphi_j}{h}, \quad 0 \leqslant j \leqslant M - m,$$

$$y_j = \psi_j, \quad M - m < j \leqslant N,$$

avec ψ_j la solution exacte du problème discret. Posons $z = mh$. La fermeture (4) s'écrit alors

$$L_z y = f_z(x), \text{ où, lorsque } 0 \leqslant z < 1,$$

$$L_z y = \begin{cases} y' - \alpha(x)\, y & \text{pour } 0 < x < z, \\ y'' - p(x)\, y & \text{pour } z \leqslant x < 1, \\ y & \text{pour } x = 1, \end{cases}$$

$$f_z(x) = \begin{cases} \beta(x) & \text{pour } 0 < x < z, \\ f(x) & \text{pour } z \leqslant x < 1, \\ b & \text{pour } x = 1 \end{cases}$$

et

$$L_z y = \begin{cases} y' - \alpha(x)\, y & \text{pour } 0 < x < 2 - z, \\ y & \text{pour } 2 - z \leqslant x \leqslant 1, \end{cases}$$

$$f_z(x) = \begin{cases} \beta(x) & \text{pour } 0 < x < 2 - z, \\ \psi(x) & \text{pour } 2 - z \leqslant x \leqslant 1, \end{cases}$$

lorsque $1 \leqslant z \leqslant 2$; $\psi(x)$ est la solution exacte du problème différentiel (on a omis $L_z y|_{x=0} = a$).

Voyons comment se comportent les fonctions $\alpha(x)$ et $\beta(x)$. Soit $p(x) > 0$; alors, conformément à $\alpha' = p(x) - \alpha^2$,

$$\alpha' < 0 \quad \text{pour} \quad \alpha > D_1 = \sqrt{\max_{[0, X]} p(x)}$$

et

$$\alpha' > 0 \quad \text{pour} \quad \alpha < D_2 = \sqrt{\min_{[0, X]} p(x)}.$$

Le champ de courbes intégrales correspondant est représenté fig. 9.4.1. Aussi la solution $\alpha(x)$ qui vérifie la condition $\alpha(x) \sim x^{-1}$ lorsque $x \to 0$, décroît de facon monotone, au moins tant qu'elle ne tombe pas dans le domaine $D_2 < \alpha < D_1$ dans lequel elle reste. L'équation (14) est linéaire en $\beta(x)$. Comme le coefficient $\alpha(x)$ est fini lorsque $x \neq 0$, la solution de cette équation reste bornée pour tous les $x \neq 0$. Par conséquent, on serait pleinement autorisé à parler d'une fermeture régulière de l'algorithme de résolution du problème proposé n'était-ce le caractère non borné de $\alpha(x)$ et $\beta(x)$ avec x tendant vers zéro. Dans le cas examiné, les valeurs α_0 et β_0 au point $n = 0$ sont des quantités de l'ordre de h^{-1}. Elles tendent à décroître avec la croissance de n par suite d'une tendance similaire des solutions d'équations différentielles. Nous savons déjà que dans nombre de cas une erreur de calcul en $2^{-t}h^{-p}$ est inévitable et admissible, et il se peut que ce caractère non régulier de l'algorithme ne détermine pas une influence prohibitive de l'erreur de calcul.

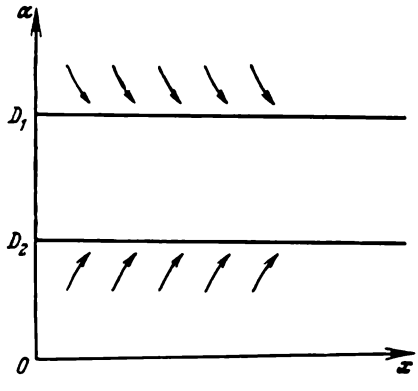


Fig. 9.4.1

Etudions maintenant le cas où $p(x)$ n'est pas partout $\geqslant 0$. Alors, il est possible que $W^1(y) = 0$ en un point intérieur du segment et, par conséquent, $\alpha(y) = \infty$. La fermeture de l'algorithme est alors irrégulière, encore qu'il ne faille pas renoncer définitivement à la méthode du balayage. Si $W^1(nh)\, h^{-2} \gg 1$ en tous les nœuds du réseau, on a, selon (16)-(18),

$$\max_n |\alpha_n| = O(h^{-2}). \tag{20}$$

Une telle irrégularité de l'algorithme n'est pas nécessairement, nous le répétons, une catastrophe. La fonction $W^1(x)$ est régulière avec la dérivée non nulle en des points y tels que $W^1(y) = 0$. C'est pourquoi la valeur $\min_{n>0} W^1(nh)$ est fréquemment une quantité en h et la relation (20) est juste. On doit donc s'attendre que, pour $p(x) \leqslant 0$ aussi, les résultats de la méthode du balayage soient insensibles à toute sorte de perturbations.

Il se trouve que la fermeture de l'algorithme de balayage est régulière si $W^1(x) \neq 0$ lorsque $x \subset (0, X]$. La condition $W'(x_0) \neq 0$ est équivalente à la condition de possibilité du problème aux limites

$$y(0) = a, \quad y(x_0) = b, \quad y'' - p(x)\, y = f(x) \qquad (21)$$

sur $[0, x_0]$ quels que soient a, b et $f(x)$. La condition de régularité de la fermeture de l'algorithme de balayage s'énonce donc comme suit :

La fermeture de la méthode du balayage est régulière si, quel que soit le segment $[0, x_0]$, $0 < x_0 \leqslant X$, *le problème aux limites* (2) *admet une solution.*

Démontrons cette assertion d'une autre manière. Revenons aux formules de remontée de la méthode du balayage. Les coefficients C_j et φ_j, $j < n$, dépendent seulement de a et des valeurs p_j et f_j pour $j < n$, i.e. aux points $jh \subset [0, nh)$. Par conséquent, dès qu'on trouve la valeur y_n, la marche des calculs est la même que dans le cas du problème aux limites pour l'équation $y'' - p(x)\, y = f(x)$ sur $[0, nh]$ avec $y(0) = a$ et $y(nh) = y_n$ donnés. Si ce problème aux limites ne possède pas de solution pour n'importe quels seconds membres $f(x)$, il y a peu de chances que le problème discret correspondant se prête bien à une résolution numérique. Aussi, si, pour un certain $x_0 \subset (0, X]$, le problème aux limites (21) est résoluble sous certaines conditions imposées à ses seconds membres, on se doute que l'algorithme de calcul est sensible à l'erreur d'arrondi, ce qui implique une étude plus méticuleuse.

Notons que la régularité de la fermeture ne garantit pas la petitesse de l'erreur de calcul totale. En effet, celle-ci est fonction des erreurs commises au cours du calcul et des facteurs de proportionnalité qui les précèdent dans l'expression de l'erreur totale. Or, la régularité de la fermeture ne garantit que la petitesse des erreurs d'arrondi engendrées. Afin que les facteurs de proportionnalité ne soient pas élevés, la solution des équations de la fermeture doit être peu sensible aux perturbations des coefficients. Et même cela ne suffit en général pas.

L'exemple de l'équation $y' = My$, $M < 0$, nous a montré que la solution du problème différentiel peut être peu sensible à de telles perturbations, tandis que celle du problème discret l'est fortement.

§ 5. Position des problèmes aux limites pour des systèmes linéaires du premier ordre

Soit le problème aux limites

$$\mathbf{y}' - A(x)\, \mathbf{y} = \mathbf{f}(x), \quad B\mathbf{y}(0) = \mathbf{b}, \quad D\mathbf{y}(X) = \mathbf{d}, \qquad (1)$$

$\mathbf{y}$, $\mathbf{f}$, $\mathbf{b}$, $\mathbf{d}$ étant des vecteurs de dimension l, l, $l - r$, r respectivement, et A, B, D des $l \times l$, $(l - r) \times l$, $r \times l$-matrices. On suppose dans la suite de l'exposé que le rang de la matrice B est $l - r$ et celui de la matrice D est r.

Avant de passer à la recherche des méthodes de résolution acceptables, discutons le degré de stabilité des solutions des problèmes aux limites par rapport à diverses perturbations et prenons pour modèle le problème aux limites

$$\mathbf{y}' - A\mathbf{y} = 0, \quad A = \text{const}, \quad B\mathbf{y}(0) = \mathbf{b}, \quad D\mathbf{y}(X) = \mathbf{d}. \qquad (2)$$

Examinons le cas d'une matrice A dont toutes les valeurs propres sont distinctes; dans le cas général, les raisonnements ne varient guère. Soient $\lambda_j = \alpha_j + i\beta_j$ les valeurs propres de A écrites dans l'ordre de croissance de α_j et $\mathbf{e}_j$ les vecteurs propres associés, $\|\mathbf{e}_j\| = 1$. La solution générale du système $\mathbf{y}' - A\mathbf{y} = 0$ s'écrit

$$\mathbf{y}(x) = \sum_{j=1}^{l} c_j \mathbf{e}_j \exp(\lambda_j x). \qquad (3)$$

Partageons λ_j en trois groupes en affectant les valeurs propres de chacun d'eux d'un indice supérieur correspondant. Les valeurs propres λ_j telles que $\exp(|\alpha_j| X)$ est très grand seront notées λ_j^+ si $\alpha_j > 0$ et λ_j^- si $\alpha_j < 0$; les valeurs propres restantes, i. e. pour lesquelles $\exp(|\alpha_j| X)$ n'est pas très élevé, se noteront λ_j^0. Indiçons supérieurement de façon correspondante les vecteurs propres $\mathbf{e}_j$ et désignons par $\sum^+$, $\sum^0$, $\sum^-$ la sommation sur les indices j associés à ces groupes. Soit l^-, l^0, l^+ le nombre de valeurs propres des groupes. Avec l'écriture (3) les extrémités du segment d'intégration se trouvent dans des positions différentes: au point $x = 0$, toutes les fonctions $\exp(\lambda_j x)$ sont de l'ordre 1, alors qu'au point $x = X$, certaines d'entre elles sont très petites et d'autres très grandes. On préférera donc à cette écriture

$$y(x) = \sum^- c_j^- \mathbf{e}_j^- \exp(\lambda_j^- x) + \sum^0 c_j^0 \mathbf{e}_j^0 \exp(\lambda_j^0 x) + \sum^+ c_j^+ \mathbf{e}_j^+ \exp(\lambda_j^+ (x - X)).$$

Ecrivons le système d'équations $B\mathbf{y}(0) = \mathbf{b}$, $D\mathbf{y}(X) = \mathbf{d}$, à l'aide duquel on doit déterminer les constantes c_j correspondant à la solution cherchée:

$$\begin{aligned} B\mathbf{y}(0) &= \sum^- c_j^- B\mathbf{e}_j^- + \sum^0 c_j^0 B\mathbf{e}_j^0 + \sum^+ c_j^+ B\mathbf{e}_j^+ \exp(-\lambda_j^+ X) = \mathbf{b}, \\ D\mathbf{y}(X) &= \sum^- c_j^- B\mathbf{e}_j^- \exp(\lambda_j^- X) + \sum^0 c_j^0 B\mathbf{e}_j^0 \exp(\lambda_j^0 X) + \\ &\qquad + \sum^+ c_j^+ B\mathbf{e}_j^+ = \mathbf{d}. \end{aligned} \qquad (4)$$

Le système (4) se récrit

$$G\mathbf{c} = \mathbf{g}$$

avec

$$\mathbf{c} = (c_1^-, \ldots, c_l^+)^T, \quad \mathbf{g} = (b_1, \ldots, b_{l-r}, d_1, \ldots, d_r)^T,$$

et G une matrice partitionnée

$$G = \begin{matrix} \begin{pmatrix} G_1^- & G_1^0 & G_1^+ \\ G_2^- & G_2^0 & G_2^+ \end{pmatrix} \begin{matrix} l-r \\ r \end{matrix} \\ \begin{matrix} l^- & l^0 & l^+ \end{matrix} \end{matrix},$$

où

$$G_1^- = [(Be_j^-)_q], \quad 1 \leqslant j \leqslant l^-, \qquad 1 \leqslant q \leqslant l-r,$$
$$G_1^0 = [(Be_j^0)_q], \quad l^- < j \leqslant l^- + l^0, \quad 1 \leqslant q \leqslant l-r,$$
$$G_1^+ = [(Be_j^+)_q \exp(-\lambda_j^+ X)], \qquad l^- + l^0 < j \leqslant l, \quad 1 \leqslant q \leqslant l-r,$$
$$G_2^- = [(De_j^-)_q \exp(\lambda_j^- X)], \qquad 1 \leqslant j \leqslant l^-, \quad 1 \leqslant q \leqslant r,$$
$$G_2^0 = [(De_j^0)_q \exp(\lambda_j^0 X)], \qquad l^- < j \leqslant l^- + l^0, \quad 1 \leqslant q \leqslant r,$$
$$G_2^+ = [(De_j^+)_q], \qquad l^- + l^0 < j \leqslant l, \quad 1 \leqslant q \leqslant r.$$

Supposons que $\Delta = \det G \neq 0$ et représentons la solution du système (4) par $\mathbf{c} = G^{-1}\mathbf{g}$; selon les formules connues donnant les éléments de la matrice inverse:

$$G^{-1} = \left[\frac{(-1)^{i+j} G_{ji}}{\Delta}\right], \tag{5}$$

G_{ji} étant les mineurs de G. Avec les hypothèses admises, les quantités $\exp(-\lambda_j^+ X)$, $\exp(\lambda_j^- X)$ sont négligeables; aussi il en est de même des éléments des matrices G_1^+ et G_2^- et le déterminant

$$\Delta = \det \begin{pmatrix} G_1^- & G_1^0 & G_1^+ \\ G_2^- & G_2^0 & G_2^+ \end{pmatrix}$$

est voisin du déterminant

$$\Delta_0 = \det \begin{pmatrix} G_1^- & G_1^0 & 0 \\ 0 & G_2^0 & G_2^+ \end{pmatrix}.$$

Distinguons deux cas:

a) le déterminant Δ_0 n'est pas petit,

b) le déterminant Δ_0 est petit, et en particulier nul.

Puisque parmi les éléments de la matrice G il n'y en a pas qui sont grands, par suite de la formule (5) les éléments de la matrice G^{-1} ne sont généralement pas très importants dans le cas a).

Supposons que les seconds membres $\mathbf{b}$ et $\mathbf{d}$ des conditions aux limites contiennent certaines erreurs $\delta\mathbf{b}$ et $\delta\mathbf{d}$ et soit $\delta\mathbf{g} = (\delta\mathbf{b}, \delta\mathbf{d})^T$. Le vecteur $\mathbf{c}$ est alors faux de $G^{-1}\delta\mathbf{g}$. Si les éléments de la matrice G^{-1} ne sont pas très importants, l'influence de $\delta\mathbf{g}$ sur les coefficients c_j s'avère modérée. La solution du problème est combinaison linéaire des termes

$$\mathbf{e}_j^- \exp(\lambda_j^- x), \quad \mathbf{e}_j^0 \exp(\lambda_j^0 x), \quad \mathbf{e}_j^+ \exp(\lambda_j^+ (x - X)$$

avec les coefficients c_j. C'est pourquoi l'erreur sur la solution approchée engendrée par $\delta \mathbf{b}$ et $\delta \mathbf{d}$ est elle aussi admissible.

Une fois de plus nos raisonnements ne sont pas suffisamment rigoureux: nous disons « petit », « très petit », « pas très grand », « très grand ». En analysant son problème, l'utilisateur doit décider lui-même de l'ordre de grandeur admissible des quantités considérées. Il se peut en particulier que dans le cas d'un système d'ordre « grand », de coefficients « pas très élevés » et un déterminant Δ_0 « pas très petit », les mineurs formés de sommes des produits d'éléments en nombre notable soient « inadmissiblement grands ».

Lorsque Δ_0 est très petit ou égal à zéro, par suite de

$$\det G \det G^{-1} = 1,$$

certains éléments de la matrice G^{-1} sont grands. Des perturbations faibles des seconds membres des conditions aux limites sont alors susceptibles de perturber notablement les coefficients c_j et donc la solution du problème.

Connaissant les matrices G^{-1} et les valeurs propres de la matrice A, les relations ci-dessus fournissent une information assez fidèle sur la perturbation de la solution du problème différentiel. Cette information ne s'obtient cependant qu'au bout de longs calculs qui se prolongent encore si on effectue les mêmes constructions dans le cas d'une matrice $A(x)$ variable. Nous allons donc essayer d'obtenir des critères de stabilité de la solution par rapport aux perturbations $\delta \mathbf{b}$ et $\delta \mathbf{d}$ qui nécessitent une information moindre quitte à être moins sûrs. Cela peut être par exemple les relations entre l^-, l^0, l^+, $l - r$ et r. Les éléments non nuls des $l - r$ premières lignes de la matrice G^0 peuvent se trouver dans les $l^- + l^0$ premières colonnes correspondant aux matrices G_1^-, G_1^0. Si $l^+ > r$, alors $l^- + l^0 < l - r$ et tous les mineurs d'ordre $(l - r) \times (l - r)$ des $l - r$ premières lignes deviennent nuls. Le développement du déterminant Δ_0 suivant les $l - r$ premières lignes donne $\Delta_0 = 0$. De même, si $l^- > l - r$, alors tous les mineurs d'ordre $r \times r$ des r dernières lignes s'annulent et donc $\Delta_0 = 0$. Lorsque $l^+ \leqslant r$ et $l^- \leqslant l - r$, le déterminant Δ_0 est combinaison linéaire de produits d'éléments des matrices B et D et de coordonnées des vecteurs propres $\mathbf{e}_j$, avec pour coefficients de ces produits les produits des nombres $\exp(\lambda_j^0 X)$ ayant un module ni très grand ni très petit. On peut admettre que ce déterminant n'est petit qu'assez rarement. La solution du problème (2) est alors peu sensible aux perturbations $\delta \mathbf{b}$ et $\delta \mathbf{d}$ subies par les seconds membres des conditions aux limites.

Forts de conclusions obtenues nous énonçons la proposition suivante. *Si $l^+ > r$ ou $l^- > l - r$, alors la solution du problème différentiel est très sensible aux perturbations des seconds membres des conditions aux limites; si $l^+ \leqslant r$ et $l^- \leqslant l - r$, alors, d'une façon générale, la solution du problème* (2) *est peu sensible aux variations des seconds membres des conditions aux limites.*

Voici un autre énoncé de la première partie de cette proposition : *pour que le problème soit peu sensible aux perturbations des conditions aux limites, il faut que le nombre des solutions particulières indépendantes* $\mathbf{e}_j \exp(\lambda_j x)$ *qui croissent fortement sur* $[0, X]$ *avec* x *soit au plus égal à celui des conditions aux limites à l'extrémité droite et le nombre des* $\mathbf{e}_j \exp(\lambda_j x)$ *décroissant fortement sur* $[0, X]$ *avec la croissance de* x *au plus égal à celui des conditions aux limites à l'extrémité gauche.*

Moyennant certaines précisions cette formulation s'étend au cas du problème (1) avec une matrice $A(x)$ variable. Une reformulation rigoureuse de cette proposition est une opération assez encombrante. Pourtant, si $A(x)$ varie de façon relativement régulière sur $[0, X]$, on se borne souvent, aux étapes initiales de l'étude de la stabilité du problème par rapport aux perturbations des conditions aux limites, à calculer le nombre de valeurs propres de $A(x)$ avec $X \operatorname{Re} \lambda_j(x)$ positif et négatif important.

Un problème aux limites est par définition *bien conditionné* (*bien posé*) si de faibles perturbations des coefficients et seconds membres des équations et des conditions aux limites entraînent une variation du même ordre de la solution du problème. Et voici une définition plus rigoureuse. On considère, en plus du problème aux limites (1), les problèmes

$$\tilde{\mathbf{y}}' - (A(x) + \delta A(x))\,\tilde{\mathbf{y}} = \mathbf{f}(x) + \delta \mathbf{f}(x),$$

$$(B + \delta B)\,\tilde{\mathbf{y}}(0) = \mathbf{b} + \delta \mathbf{b}, \quad (D + \delta D)\,\tilde{\mathbf{y}}(X) = \mathbf{d} + \delta \mathbf{d}$$

avec une mesure de la perturbation pas très grande

$$\varepsilon = \max_{0 \leqslant x \leqslant X} (\| \delta A(x) \| + \| \delta \mathbf{f}(x) \|) + \| \delta B \| + \| \delta D \| + \| \delta \mathbf{b} \| + \| \delta \mathbf{d} \|.$$

Si toutes les solutions de tels problèmes aux limites vérifient les inégalités

$$\max_{0 \leqslant x \leqslant X} \| \tilde{\mathbf{y}}(x) - \mathbf{y}(x) \| \leqslant M\varepsilon \tag{6}$$

avec M constant pas très élevé, on dit du problème initial qu'il est *bien conditionné* ; dans le cas contraire, ce problème est *mal conditionné*. La valeur minimale $M(\varepsilon_0)$ pour laquelle l'inégalité (6) est juste quel que soit $\varepsilon \leqslant \varepsilon_0$ ($\varepsilon_0 > 0$ fixe) est parfois appelée le *nombre de conditionnement* du problème donné (relativement à des perturbations de norme ε_0 au plus). Le conditionnement d'un problème caractérise la stabilité de la solution par rapport aux perturbations des valeurs initiales, par exemple à une donnée imprécise des paramètres figurant dans les coefficients de l'équation. Comme les erreurs d'arrondi sont équivalentes aux perturbations des coefficients de l'équation de départ, le nombre de conditionnement caractérise aussi la stabilité de la solution numérique vis-à-vis à toutes sortes d'arrondis. Si on connaît une estimation ε approximative de la perturbation des coefficients du problème et si une erreur en $M(\varepsilon)\,\varepsilon$

est permise, il vaut la peine de procéder à une résolution numérique directe.

Considérons en guise d'exemple le problème de Cauchy pour le système $y_1' = y_2$, $y_2' = y_1$ lorsque $y_1(0) = 1$, $y_2(0) = 1$ sur le segment $[0, 30]$. Les valeurs propres de la matrice $A = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$ valent ± 1. $\exp(-30)$ est très petite et $\exp(30)$ très grande, $l^- = 1$, $l^0 = 0$, $l^+ = 1$, $l - r = 2$, $r = 0$ et, par conséquent, $l^- = 1 > r = 0$. Les perturbations insignifiantes de la condition aux limites conduisent à des variations fort importantes de la solution. Dans le cas considéré les constructions du présent paragraphe ne rendent qu'un piètre service car on comprend de toute façon qu'une croissance brusque des solutions de l'équation linéaire initiale entraîne une croissance rapide de l'erreur sur la solution.

Examinons, pour ce même système, un problème avec les conditions aux limites $b_1 y_1(0) + b_2 y_2(0) = b$, $d_1 y_1(30) + d_2 y_2(30) = d$; les vecteurs propres associés aux valeurs propres $\lambda_1^- = -1$ et $\lambda_2^+ = 1$ sont respectivement $(1, -1)^T$ et $(1, 1)^T$, $l^+ = r$, $l^- = l - r$. En principe, le problème devrait être stable par rapport aux perturbations δb et δd. La solution est recherchée sous la forme

$$c_1^- \begin{pmatrix} 1 \\ -1 \end{pmatrix} \exp(-x) + c_2^+ \begin{pmatrix} 1 \\ 1 \end{pmatrix} \exp(x - 30).$$

Les équations (4) s'écrivent

$$\begin{aligned} (b_1 - b_2)\, c_1^- + (b_1 + b_2)\, c_2^+ \exp(-30) &= b, \\ (d_1 - d_2)\, c_1^- \exp(-30) + (d_1 + d_2)\, c_2^+ &= d. \end{aligned} \tag{7}$$

Il en résulte

$$c_1^- = \frac{b(d_1 + d_2) - d(b_1 + b_2)\exp(-30)}{\Delta},$$

$$c_2^+ = \frac{d(b_1 - b_2) - b(d_1 - d_2)\exp(-30)}{\Delta},$$

avec

$$\Delta = (b_1 - b_2)(d_1 + d_2) - (d_1 - d_2)(b_1 + b_2)\exp(-60)$$

voisin de $\Delta_0 = (b_1 - b_2)(d_1 + d_2)$. Pour des coefficients b_i et d_i faibles et Δ_0 grand, les coefficients c_1^- et c_2^+ changent peu pour de petites variations δb et δd. Lorsque $\Delta = 0$, le système (7) admet une solution non nulle c_1^-, c_2^+ pour $b = d = 0$ et on dit que le problème « se trouve sur le spectre ». Cela signifie que le problème homogène possède une solution non nulle et qu'il est inutile de parler de la stabilité de la solution par rapport aux perturbations des conditions aux limites. Si $\Delta_0 = 0$, le problème est mal conditionné.

Aux paragraphes 6-8 le problème aux limites proposé est supposé bien conditionné.

PROBLÈME. Démontrer qu'un problème aux limites bien conditionné possède une solution unique.

§ 6. Algorithmes de résolution des problèmes aux limites pour des systèmes d'équations du premier ordre

La *méthode du tir* est la technique la plus simple de résolution du problème aux limites (5.1). Soit le système

$$B\mathbf{y}(0) = \mathbf{b}. \tag{1}$$

Puisque le rang de la matrice B est par hypothèse égal à $l - r$, la solution générale de (1) s'écrit

$$\mathbf{y}_0 + \sum_{i=1}^{r} c_i \mathbf{y}_i,$$

où $\mathbf{y}_0$ est une solution arbitraire du système non homogène $B\mathbf{y} = \mathbf{b}$ et $\mathbf{y}_1, \ldots, \mathbf{y}_r$ un système de r solutions linéairement indépendantes de $B\mathbf{y} = 0$. Soit $\mathbf{y}_1, \ldots, \mathbf{y}_r, \mathbf{y}_0$ un certain jeu de tels vecteurs. Trouvons par intégration numérique une solution particulière du système non homogène

$$\mathbf{y}_0' = A(x)\,\mathbf{y}_0 + \mathbf{f}(x) \tag{2}$$

pour la condition initiale $\mathbf{y}_0(0) = \mathbf{y}_0$ et les solutions du système homogène

$$\mathbf{y}_j' = A(x)\,\mathbf{y}_j, \quad j = 1, \ldots, r \tag{3}$$

avec les conditions initiales $\mathbf{y}_j(0) = \mathbf{y}_j$. Supposons que $\mathbf{y}(x)$ satisfait à (2) et à la condition à l'extrémité gauche (1). Mettons le vecteur $\mathbf{y}(0)$ sous la forme

$$\mathbf{y}(0) = \mathbf{y}_0 + \sum_{j=1}^{r} c_j \mathbf{y}_j;$$

la fonction vectorielle

$$\mathbf{z}(x) = \mathbf{y}_0(x) + \sum_{j=1}^{r} c_j \mathbf{y}_j(x)$$

vérifie l'équation (2) et coïncide avec $\mathbf{y}(x)$ lorsque $x = 0$. Par conséquent,

$$\mathbf{y}(x) = \mathbf{y}_0(x) + \sum_{j=1}^{r} c_j \mathbf{y}_j(x). \tag{4}$$

Toute fonction du type (4) satisfait aux relations (1) et (2). La variété de toutes les solutions de (2) vérifiant la condition (1) est donc donnée par l'égalité (4). La solution cherchée s'obtient en dégageant de cette variété la solution qui satisfait à la condition à l'extrémité droite. Déterminons les coefficients c_j à l'aide du système de r équations à r inconnues

$$D\left(\mathbf{y}_0(X) + \sum_{j=1}^{r} c_j \mathbf{y}_j(X)\right) = d. \tag{5}$$

En supposant que le problème (5.1) admette une solution unique, le déterminant de ce système est différent de zéro ; en effet, dans le cas contraire, on aurait trouvé que le problème aux limites homogène ($\mathbf{f} \equiv 0$, $\mathbf{b} = 0$, $\mathbf{d} = 0$) admet une solution non nulle

$$\mathbf{y}(x) = \sum_{j=1}^{r} c_j \mathbf{y}_j(x).$$

Si $c_1, \ldots, c_r$ est solution du système (5), la fonction vectorielle

$$\mathbf{y}(x) = \mathbf{y}_0(x) + \sum_{j=1}^{r} c_j \mathbf{y}_j(x)$$

vérifie l'équation et toutes les conditions aux limites et est donc solution du problème proposé.

S'il est nécessaire d'économiser la mémoire de la machine, on trouve

$$\mathbf{y}(0) = \mathbf{y}_0(0) + \sum_{j=1}^{r} c_j \mathbf{y}_j(0)$$

ou

$$\mathbf{y}(X) = \mathbf{y}_0(X) + \sum_{j=1}^{r} c_j \mathbf{y}_j(X),$$

puis on résout numériquement, en avant ou en arrière, le problème de Cauchy.

Si certaines solutions du système homogène $\mathbf{y}' = A(x)\mathbf{y}$ croissent brusquement avec x, l'erreur de calcul sur $\mathbf{y}_0(x), \ldots, \mathbf{y}_r(x)$ est susceptible d'une augmentation aussi rapide, et la méthode du tir est en défaut. Il existe une autre explication de la déficience du procédé en question. Appliquons la méthode du tir pour $A = \text{const}$ et désignons par $\mathbf{e}_j$ et λ_j les vecteurs propres et les valeurs propres associées de la matrice A avec

$$\operatorname{Re} \lambda_1 \leqslant \ldots \leqslant \operatorname{Re} \lambda_l ;$$

supposons que

$$\mathbf{y}_j = \sum_{k=1}^{l} \alpha_{jk} \mathbf{e}_k ;$$

alors

$$\mathbf{y}_j(X) = \sum_{k=1}^{l} \alpha_{jk} \mathbf{e}_k \exp(\lambda_k X). \tag{6}$$

Si $\alpha_{jl} \neq 0$ et $\exp(\operatorname{Re} \lambda_l X) \gg \exp(\operatorname{Re} \lambda_{l-1} X)$, on a

$$\mathbf{y}_j(X) \sim \alpha_{jl} \mathbf{e}_l \exp(\lambda_l X).$$

Ainsi, toutes les lignes du déterminant du système (5) sont approximativement proportionnelles au vecteur $\mathbf{e}_l$ et la solution $c_1, \ldots, c_r$ de ce système sera affectée d'une erreur de calcul notable. Bien qu'ils soient par endroits peu rigoureux, ces raisonnements expliquent de façon vraisemblable l'éventualité d'une grande erreur de calcul dans la méthode du tir.

Proposons-nous pour le moment de résoudre un problème particulier : le calcul de $\mathbf{y}(X)$; on peut l'interpréter comme un problème d'extraction consistant à trouver parmi les vecteurs de la variété

$$\mathbf{y}_0(X) + \sum_{j=1}^{r} c_j \mathbf{y}_j(X) \tag{7}$$

celui qui vérifie la condition à l'extrémité droite $D\mathbf{y}(X) = \mathbf{d}_0$. En le résolvant directement, on risque de commettre, nous l'avons vu, une erreur importante. Les vecteurs $\mathbf{y}_j(X)$ seront proches de vecteurs proportionnels ou, comme on le dit quelquefois, la base des vecteurs $\mathbf{y}_j(x)$ « s'aplatira » lorsqu'on se rapprochera du point X.

Cependant, il n'est nullement besoin de se donner la variété (7) à chaque instant au moyen du vecteur $\mathbf{y}_0(x)$ et de la base $\{\mathbf{y}_j(x)\}$. A mesure que la base « s'aplatit », il y a lieu de se donner de temps en temps la variété (7) d'une autre façon. C'est là l'idée de la première méthode du balayage orthogonal.

Partageons le segment $[0, X]$ par les points $0 = X_0 < X_1 < \ldots < X_m = X$. Définissons comme suit l'opération M_s sur la $l \times (r+1)$-matrice $G = (\mathbf{g}_1, \ldots, \mathbf{g}_r, \mathbf{g}_0)$ en tant qu'une opération effectuée sur ses vecteurs colonnes $\mathbf{g}_j$.

Soient $\mathbf{g}_j^s(x)$, $j = 1, \ldots, r$, la solution du système $\mathbf{y}' = A(x)\mathbf{y}$ avec la condition initiale $\mathbf{y}(X_s) = \mathbf{g}_j$ et $\mathbf{g}_0^s(x)$ celle du système $\mathbf{y}' = A(x)\mathbf{y} + \mathbf{f}(x)$ avec la condition initiale $\mathbf{y}(X_s) = \mathbf{g}_0$. On a alors $M_s\mathbf{g}_j = \mathbf{g}_j^s(X_{s+1})$.

Définissons l'opération d'orthogonalisation U des colonnes de la matrice G. Notons respectivement $\mathbf{q}_1, \ldots, \mathbf{q}_r, \mathbf{q}_0$ les colonnes de $Q = U(G)$. Les vecteurs $\mathbf{q}_1, \ldots, \mathbf{q}_r, \mathbf{q}_0$ s'obtiennent des $\mathbf{g}_1, \ldots, \mathbf{g}_r, \mathbf{g}_0$ par orthogonalisation successive et par normalisation, le vecteur $\mathbf{q}_0$ n'étant pas normé. Ainsi, on obtient successivement (comme au § 15, chap. II) les vecteurs

$$\mathbf{z}_j = \mathbf{g}_j - \sum_{i=1}^{j-1} (\mathbf{g}_j, \mathbf{q}_i)\mathbf{q}_i, \tag{8}$$

$$\mathbf{q}_j = \frac{\mathbf{z}_j}{\sqrt{(\mathbf{z}_j, \mathbf{z}_j)}}, \quad j = 1, \ldots, r,$$

$$\mathbf{q}_0 = \mathbf{g}_0 - \sum_{i=1}^{r} (\mathbf{g}_0, \mathbf{q}_i)\mathbf{q}_i. \tag{9}$$

Exprimons les z_j des relations (8), (9) moyennant q_i et portons tous les q_i dans les premiers membres : il vient

$$\sqrt{(z_j, z_j)}\, q_j + \sum_{i=1}^{j-1} (g_j, q_i)\, q_i = g_j, \quad j = 1, \ldots, r,$$

$$\sum_{i=1}^{r} (g_0, q_i)\, q_i + q_0 = g_0.$$

Ce système s'écrit sous forme matricielle $Q\Omega = G$, où

$$\Omega = \begin{pmatrix} \omega_{11} & \omega_{21} & \ldots & \omega_{r1} & \omega_{r+1,1} \\ 0 & \omega_{22} & \ldots & \omega_{r2} & \omega_{r+1,2} \\ \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & \ldots & \omega_{rr} & \omega_{r+1,r} \\ 0 & 0 & \ldots & 0 & 1 \end{pmatrix},$$

$\omega_{jj} = \sqrt{(z_j, z_j)}$, $\omega_{ji} = (g_j, q_i)$ lorsque $i < j \leqslant r$, $\omega_{r+1,i} = (g_0, q_i)$; autrement dit, $Q = U(G) = G \cdot \Omega^{-1}$. Choisissons les vecteurs $y_1, \ldots, y_r, y_0$ comme dans la méthode du tir et orthogonalisons la matrice $Y_0 = (y_1, \ldots, y_r, y_0)$. Soit

$$G_0(X_0) = U(Y_0) = (g_1^0(X_0), \ldots, g_r^0(X_0), g_0^0(X_0)).$$

L'opération M_0 sur $G_0(X_0)$ donne $G_0(X_1) = M_0(G_0(X_0))$; il vient ensuite la matrice

$$G_1(X_1) = U(G_0(X_1)) = G_0(X_1)\,\Omega_1^{-1} \text{ et ainsi de suite.} \qquad (10)$$

Désignons par $G_s(x)$ les matrices $(g_1^s(x), \ldots, g_r^s(x), g_0^s(x))$. Les calculs conduisent finalement aux matrices

$$\begin{aligned} &G_0(X_0),\ G_0(X_1) = M_0(G_0(X_0)), \\ &\quad G_1(X_1) = U(G_0(X_1)) = G_0(X_1)\,\Omega_1^{-1}, \\ &\quad G_1(X_2) = M_1(G_1(X_1)), \ \ldots, \\ &G_{m-1}(X_{m-1}) = U(G_{m-2}(X_{m-1})) = G_{m-2}(X_{m-1})\,\Omega_{m-1}^{-1}, \\ &\quad G_{m-1}(X_m) = M_{m-1}(G_{m-1}(X_{m-1})), \\ &\quad G_m(X_m) = U(G_{m-1}(X_m)) = G_{m-1}(X_m)\,\Omega_m^{-1}. \end{aligned} \qquad (11)$$

LEMME. *La variété balayée par les extrémités des vecteurs*

$$y = g_0^s(X_s) + \sum_{j=1}^{r} c_j g_j^s(X_s) = G_s(X_s)\, c \qquad (12)$$

avec $c_1, \ldots, c_r$ quelconques engendre la variété L_s de toutes les valeurs $y(X_s)$ des solutions du système $y' = A(x)\,y + f(x)$ satisfaisant à la condition à l'extrémité gauche $By(0) = b$; ici

$$c = (c_1, \ldots, c_r, 1)^T.$$

DÉMONSTRATION. Le lemme est évident pour $s=0$. Supposons qu'il en est de même pour un certain s. Avec l'opération M_s nous passons de la variété L_s à la variété L_{s+1} des valeurs de telles solutions au point x_{s+1} :

$$\mathbf{y}=G_s(X_{s+1})\,\mathbf{c}.$$

Comme

$$G_s(X_{s+1})=G_{s+1}(X_{s+1})\,\Omega_{s+1},$$

les éléments de cette variété s'écrivent

$$\mathbf{y}=G_{s+1}(X_{s+1})\,\Omega_{s+1}\mathbf{c}.$$

Posons $\Omega_{s+1}\mathbf{c}=\tilde{\mathbf{c}}$; du moment que $\omega^{s+1}_{r+1,\,r+1}=1$, la dernière composante du vecteur $\tilde{\mathbf{c}}$ vaut 1 et en changeant de notation $\tilde{\mathbf{c}}=\mathbf{c}$, nous obtenons une écriture des éléments de L_{s+1} analogue à (12), à savoir $\mathbf{y}=G_{s+1}(X_{s+1})\,\mathbf{c}$. C.Q.F.D.

Le vecteur $\mathbf{g}^s_0(x_s)$ est orthogonal par construction à tous les vecteurs $\mathbf{g}^s_j(x_s)$, $j=1,\ \ldots,\ r$, constituant une base de la variété L_s. Sa norme est donc inférieure à celle de tout vecteur élément de L_s. Supposons que sur chacun des segments $[X_s,\ X_{s+1}]$ la solution cherchée du problème s'écrit

$$\mathbf{y}(x)=\mathbf{g}^s_0(x)+\sum_{j=1}^{r}\beta^s_j\mathbf{g}^s_j(x),$$

ou encore

$$\mathbf{y}(x)=G_s(x)\,\boldsymbol{\beta}^s, \tag{13}$$

où $\boldsymbol{\beta}^s=(\beta^s_1,\ \ldots,\ \beta^s_r,\ 1)$.

Faisant dans (13) $s=q$, $x=X_{q+1}$, nous trouvons

$$\mathbf{y}(X_{q+1})=G_q(X_{q+1})\,\boldsymbol{\beta}^q,$$

et faisant $s=q+1$, $x=X_{q+1}$

$$\mathbf{y}(X_{q+1})=G_{q+1}(X_{q+1})\,\boldsymbol{\beta}^{q+1}.$$

D'autre part,

$$(G_{q+1}(X_{q+1}))^{-1}G_q(X_{q+1})=\Omega_{q+1}.$$

Les trois dernières égalités entraînent

$$\boldsymbol{\beta}^{q+1}=\Omega_{q+1}\boldsymbol{\beta}^q. \tag{14}$$

La valeur $\boldsymbol{\beta}^m$ s'obtient à partir de la condition aux limites

$$D\mathbf{y}(X)=D(G_m(X_m)\,\boldsymbol{\beta}^m)=\mathbf{d}.$$

Ensuite, à l'aide de l'égalité (14), $q=m-1,\ \ldots,\ 0$, nous définissons de proche en proche $\boldsymbol{\beta}^{m-1},\ \ldots,\ \boldsymbol{\beta}^0$.

En restant sur le plan formel nous devrions maintenant recourir à la formule (13) pour déterminer $\mathbf{y}(x)$ sur $[X_s,\ X_{s+1}]$. Cette façon

d'agir conduirait cependant à une occupation mémoire considérable vu qu'il faut mettre en réserve les valeurs de $G_s(x)$ en de nombreux points. Pour éviter cet inconvénient, il vaut donc mieux calculer $y(X_s)$, puis résoudre les problèmes de Cauchy sur chaque $[X_s, X_{s+1}]$.

Au lieu d'une étude complète de l'influence de l'erreur de calcul dans cet algorithme nous nous contenterons de certaines considérations permettant d'affirmer qu'il ne devrait pas être très sensible aux arrondis si les segments partiels $[X_s, X_{s+1}]$ ne sont pas très grands.

Puisque la norme de $g_0^s(X_s)$ est inférieure à celle de tout élément de la variété L_s, on a, en particulier,

$$\| g_0^s(X_s) \| \leqslant \| y(X_s) \| \leqslant Y = \max_{0 \leqslant x \leqslant X} \| y(x) \|,$$

où $y(x)$ est la solution cherchée. Soit maintenant $X_{k+1} \to X_k$, X_k fixe. La matrice

$$G_k(X_{k+1}) = (g_1^k(X_{k+1}), \ldots, g_r^k(X_{k+1}), g_0^k(X_{k+1}))$$

tend alors vers

$$G_k(X_k) = (g_1^k(X_k), \ldots, g_r^k(X_k), g_0^k(X_k)).$$

Pour cette dernière matrice, la transformation $U(G_k(X_k))$ qui la change en une matrice dont les r premières colonnes sont orthonormées et la dernière orthogonale aux premières est identique puisque $G_k(X_k)$ est déjà de cette forme. Aussi $\Omega_{k+1} = E$ si $X_{k+1} = X_k$. Naturellement Ω_{k+1} tend vers E lorsque $X_{k+1} \to X_k$. Les vecteurs des r premières colonnes de la matrice $G_k(X_k)$ sont munis de la norme unité et

$$\| g_0^k(X_k) \| \leqslant Y.$$

Ces vecteurs étant uniformément bornés, $G_{k+1}(X_k)$ tend vers $G_k(X_k)$ (et donc Ω_{k+1} vers E) uniformément sur l'ensemble de tous les points $X_1, \ldots, X_m$ pour $\delta = \max_k |X_{k+1} - X_k|$ tendant vers zéro, autrement dit, $\| \Omega_{k+1} - E \| \leqslant \omega(\delta)$, où $\omega(\delta)$ tend vers 0 avec δ. C'est pourquoi, lorsque δ n'est pas très élevé, les normes $\| \Omega_k \|$ des matrices ainsi que les normes $\| g_j^k(x) \|$ des fonctions sur les segments $[X_k, X_{k+1}]$ sont uniformément bornées par une constante pas très grande.

Les affirmations ci-dessus n'autorisent aucune conclusion sur la qualité de l'algorithme considéré. Nous avons cependant la chance de n'avoir affaire qu'à des quantités uniformément bornées, et ce par des nombres pas très élevés. Dans le cas contraire, l'arrondi équivaudrait probablement à des perturbations considérables des coefficients du système, ce qui nous ferait douter de la légitimité de l'algorithme en question.

§ 7. Méthodes du balayage orthogonal différentiel

Pour $\delta \to 0$ construisons la fermeture de l'algorithme décrit au paragraphe précédent. Puisque les opérateurs M_s et les opérateurs d'orthogonalisation agissent sur $g_j^s(X_s)$ de façon différente selon que $j \neq 0$ ou $j = 0$, on partitionne les matrices considérées en mettant en évidence les blocs correspondant aux solutions g_0^s de l'équation avec second membre. Posons

$$G_s(X_k) = (Y_s(X_k),\ \mathbf{y}_s(X_k)),$$

où

$$Y_s(X_k) = (g_1^s(X_k),\ \ldots,\ g_r^s(X_k)),\quad \mathbf{y}_s(X_k) = g_0^s(X_k),$$

et

$$\Omega_k = \begin{pmatrix} E + \delta_{k-1} R_k & \delta_{k-1} \mathbf{r}_k \\ 0 & 1 \end{pmatrix}$$

où

$$\delta_k = X_{k+1} - X_k,\quad \mathbf{r}_k = \left(\frac{\omega_{r+1,1}^k}{\delta_k},\ \ldots,\ \frac{\omega_{r+1,r}^k}{\delta_k} \right)^T,$$

et la matrice R_k est triangulaire droite. L'introduction de nouvelles fonctions et matrices se fait de la même manière que les transformations analogues lors de la construction de la fermeture de l'algorithme de balayage du § 4. Notons $\beta_s = (\beta_1^s,\ \ldots,\ \beta_r^s)^T$, $Z_s = Y_s(X_s)$, $\mathbf{z}_s = \mathbf{y}_s(X_s)$.

Supposons qu'il existe des fonctions vectorielles et matricielles régulières

$$Z(x),\ \mathbf{z}(x),\ R(x),\ \mathbf{r}(x),\ \beta(x)$$

telles que

$$Z_s \Rightarrow Z(x),\quad \mathbf{z}_s \Rightarrow \mathbf{z}(x),\quad R_s \Rightarrow R(x),\quad \mathbf{r}_s \Rightarrow \mathbf{r}(x),\quad \beta_s \Rightarrow \beta(x);$$

$g_s \Rightarrow g(x)$ veut dire que

$$\max_{X_s} \left(\| g_s - g(X_s) \| + \left\| \frac{g_{s+1} - g_s}{\delta_s} - g'(X_s) \right\| \right) \to 0 \quad \text{pour } \delta \to 0.$$

Une multiplication par blocs permet d'écrire

$$G_s(X_{s+1}) = G_{s+1}(X_{s+1})\,\Omega_{s+1},$$

ou, ce qui revient au même,

$$(Y_s(X_{s+1}),\ \mathbf{y}_s(X_{s+1})) = (Y_{s+1}(X_{s+1}),\ \mathbf{y}_{s+1}(X_{s+1})) \begin{pmatrix} E + \delta_s R_{s+1} & \delta_s \mathbf{r}_{s+1} \\ 0 & 1 \end{pmatrix},$$

comme l'ensemble d'égalités

$$Y_s(X_{s+1}) = Y_{s+1}(X_{s+1})\,(E + \delta_s R_{s+1}), \tag{1}$$

$$\mathbf{y}_s(X_{s+1}) = Y_{s+1}(X_{s+1})\,\delta_s \mathbf{r}_{s+1} + \mathbf{y}_{s+1}(X_{s+1}). \tag{2}$$

Puisque les colonnes des matrices $Y_s(x)$ et, par conséquent, ces matrices mêmes vérifient le système $\mathbf{y}' = A(x)\,\mathbf{y}$, on a

$$Y_s(X_{s+1}) = Y_s(X_s) + \delta_s A_s Y_s(X_s) + O(\delta_s^2).$$

De même,

$$\mathbf{y}_s(X_{s+1}) = \mathbf{y}_s(X_s) + \delta_s (A_s \mathbf{y}_s(X_s) + \mathbf{f}_s) + O(\delta_s^2),$$

avec $A_s = A(X_s)$, $\mathrm{f}_s = \mathrm{f}(X_s)$. La définition de la matrice Z_s et du vecteur z_s aidant, nous mettons les égalités (1) et (2) sous la forme

$$Z_s + \delta_s A_s Z_s = Z_{s+1}(E + \delta_s R_{s+1}) + O(\delta_s^2),$$
$$\mathrm{z}_s + \delta_s (A_s \mathrm{z}_s + \mathrm{f}_s) = Z_{s+1}\delta_s \mathrm{r}_{s+1} + \mathrm{z}_{s+1} + O(\delta_s^2),$$

ou

$$\frac{Z_{s+1} - Z_s}{\delta_s} = A_s Z_s - Z_{s+1} R_{s+1} + O(\delta_s),$$

$$\frac{\mathrm{z}_{s+1} - \mathrm{z}_s}{\delta_s} = A_s \mathrm{z}_s + \mathrm{f}_s - Z_{s+1}\mathrm{r}_{s+1} + O(\delta_s).$$

Le passage à la limite lorsque $\delta \to 0$ donne les équations différentielles

$$Z' = AZ - ZR, \tag{3}$$
$$\mathrm{z}' = A\mathrm{z} + \mathrm{f} - Z\mathrm{r}. \tag{4}$$

La condition que les colonnes de $Z_m = Y_m(X_m)$ sont orthonormées s'écrit $Z_m^T Z_m = E$ et devient à la limite $Z^T(x)\,Z(x) = E$. Il en résulte

$$(Z^T)'\,Z + Z^T Z' = E' = 0;$$

en transposant (3), on obtient l'expression de $(Z^T)'$; il vient finalement

$$(Z^T A^T - R^T Z^T)\,Z + Z^T(AZ - ZR) = 0,$$

ou encore

$$R + R^T = Z^T(A + A^T)\,Z. \tag{5}$$

Le vecteur $\mathrm{z}_{s+1} = \mathrm{y}_{s+1}(x_{s+1})$ est par construction orthogonal aux vecteurs colonnes $\mathrm{g}_j^{s+1}(x_{s+1})$ de la matrice Z_{s+1}. Cette condition s'écrit $Z_{s+1}^T \mathrm{z}_{s+1} = 0$ et devient à la limite $Z^T\mathrm{z} = 0$. Dérivons-la ; il vient

$$(Z^T)'\,\mathrm{z} + Z^T\mathrm{z}' = 0,$$

ou

$$(Z^T A^T - R^T Z^T)\,\mathrm{z} + Z^T(A\mathrm{z} + \mathrm{f} - Z\mathrm{r}) = 0.$$

Compte tenu de $Z^T\mathrm{z} = 0$, $Z^T Z = E$, la dernière relation prend la forme

$$Z^T(A + A^T)\,\mathrm{z} + Z^T\mathrm{f} - \mathrm{r} = 0,$$

ou

$$\mathrm{r} = Z^T(A + A^T)\,\mathrm{z} + Z^T\mathrm{f}.$$

Portons cette expression de r dans le deuxième membre de (4) :

$$\mathrm{z}' = A\mathrm{z} + \mathrm{f} - ZZ^T(A\mathrm{z} + \mathrm{f} + A^T\mathrm{z}) = (E - ZZ^T)(A\mathrm{z} + \mathrm{f}) - ZZ^T A^T\mathrm{z}. \tag{6}$$

Il nous reste à obtenir la relation limite pour (6.14) ; décomposée en blocs, cette égalité s'écrit

$$\begin{pmatrix} E + \delta_s R_{s+1} & \delta_s \mathrm{r}_{s+1} \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \beta_s \\ 1 \end{pmatrix} = \begin{pmatrix} \beta_{s+1} \\ 1 \end{pmatrix},$$

ce qui équivaut à

$$(E + \delta_s R_{s+1})\,\beta_s + \delta_s \mathrm{r}_{s+1} = \beta_{s+1},$$

ou à

$$\frac{\beta_{s+1} - \beta_s}{\delta_s} = R_{s+1}\beta_s + \mathrm{r}_{s+1}.$$

En passant à la limite lorsque $\delta \to 0$ on trouve

$$\beta' = R\beta + r. \tag{7}$$

Regroupons toutes les relations obtenues:

$$\begin{aligned} Z' &= AZ - ZR, \\ z' &= (E - ZZ^T)(Az + f) - ZZ^T A^T z, \end{aligned} \tag{8}$$

où la matrice triangulaire supérieure R est déterminée de façon univoque par

$$R + R^T = Z^T (A + A^T) Z; \tag{9}$$

$$\beta' = R\beta + r, \quad r = Z^T (A + A^T) z + Z^T f. \tag{10}$$

Par suite de (6.12) la solution du problème s'écrit

$$y(x) = Z(x)\,\beta(x) + z(x). \tag{11}$$

Ainsi, l'algorithme de balayage orthogonal (§ 6) est fermé par l'algorithme suivant de résolution du problème initial. Dans les conditions initiales $Z(0) = Y_0(0)$, $z(0) = g_0^0(0)$ déterminons les fonctions $Z(x)$, $z(x)$ et $R(x)$ à l'aide des relations (8), (9). Déterminons ensuite $\beta(X)$ à partir de la condition

$$D(Z(X)\,\beta(X) + z(X)) = d, \tag{12}$$

intégrons l'équation (10) dans le sens des x décroissants et calculons les valeurs de $y(x)$ par la formule (11). Il est naturel de parler une fois de plus du *balayage orthogonal*. Comme $\| z_s \| \leqslant Y$ pour tous les s, la fonction limite $z(x)$ satisfait à $\| z(x) \| \leqslant Y$. Les colonnes de la matrice Z étant orthonormées, (9) entraîne l'estimation $\| R \| \leqslant \sqrt{2l}\, \| A \|$; tous les coefficients des équations obtenues sont donc bornés par des quantités pas très importantes. Par conséquent, la fermeture formelle de l'algorithme de balayage orthogonal est régulière.

Le balayage différentiel correspondant aux relations (8)-(10) n'est pas le seul procédé possible. Soit L un sous-espace vectoriel r-dimensionnel de solutions du système homogène $y' = A(x)\,y$. Désignons par $L(x)$ la trace des éléments de ce sous-espace.

LEMME. *Soient $z_1(x_0), \ldots, z_r(x_0)$ des vecteurs colonnes formant une base du sous-espace $L(x_0)$; $Q(x)$ une $r \times r$-matrice quelconque; $Z(x)$ une $l \times r$-matrice, solution du système $Z' = AZ - ZQ$ avec la condition initiale $Z(x_0) = (z_1(x_0), \ldots, z_r(x_0))$. Alors les colonnes $z_1(x), \ldots, z_r(x)$ de la matrice $Z(x)$ appartiennent au sous-espace $L(x)$.*

DÉMONSTRATION. Si z est solution du système homogène $z' = Az$ et v solution du système adjoint

$$v' = -A^T v, \tag{13}$$

alors

$$\begin{aligned} (v, z)' &= (v', z) + (v, z') = \\ &= (-A^T v, z) + (v, Az) = -(v, Az) + (v, Az) = 0. \end{aligned} \tag{14}$$

Choisissons une base $v_1(x_0), \ldots, v_{l-r}(x_0)$ dans le supplémentaire orthogonal $L^T(x_0)$ du sous-espace $L(x_0)$ et notons $v_i(x)$ les solutions du système adjoint (13) avec les conditions initiales $v_i(x_0)$ et $V(x)$ la matrice $(v_1(x), \ldots, v_{l-r}(x))$. Si le vecteur $v_i(x_0)$ est orthogonal à tous les vecteurs du sous-espace $L(x_0)$, l'égalité (14) entraîne l'orthogonalité de $v_i(x)$ à tous les vecteurs du sous-espace $L(x)$. Il en découle que les colonnes de la matrice $V(x)$ forment une base de $L^T(x)$. Il résulte de (13) que $V' = -A^T V$ et donc $(V^T)' = -V^T A$. Portant dans $(V^T Z)' = V^T Z' + (V^T)' Z$ les valeurs $Z' = AZ - ZQ$ et $(V^T)' = -V^T A$

il vient

$$(V^TZ)' = -(V^TZ)\,Q. \tag{15}$$

$U = V^TZ$ est une $(l-r)\times r$-matrice; évidemment $u_{ij} = (\mathbf{v}_i, \mathbf{z}_j)$. Du moment que $\mathbf{z}_j(x_0)$ appartiennent à $L(x_0)$ et $\mathbf{v}_i(x_0)$ au supplémentaire orthogonal $L^T(x_0)$, $u_{ij}(x_0) = 0$ quels que soient i, j. Selon (15), les vecteurs lignes $\mathbf{u}_i(x)$ de la matrice U vérifient le système homogène d'équations linéaires $\mathbf{u}_i' = \mathbf{u}_iQ$. Dans la condition initiale $\mathbf{u}_i(x_0) = 0$ ce système admet une solution unique $\mathbf{u}_i(x) \equiv 0$. Cela signifie que $(\mathbf{v}_i(x), \mathbf{z}_j(x)) = 0$ pour tous les i, j; les vecteurs $\mathbf{z}_j(x)$ sont orthogonaux aux vecteurs $\mathbf{v}_i(x)$ qui forment une base de $L^T(x)$, et appartiennent par conséquent à $L(x)$. Le lemme est démontré.

La donnée d'une matrice $Q(x)$ arbitraire et la résolution du système (12) donnent la matrice $Z(x)$ dont les colonnes, pour tout x, appartiennent au sous-espace $L(x)$.

En discutant la méthode du tir nous avons vu que lorsque $Q(x) \equiv 0$ et pour certains $A(x)$, nous n'avons pas agi au mieux en nous donnant le sous-espace $L(x)$ par les colonnes de la matrice $Z(x)$. Aussi on prêtera une attention particulière à la façon dont on doit choisir la matrice $Q(x)$. En construisant la fermeture de l'algorithme de balayage orthogonal nous avons obtenu l'un des procédés possibles de choix de $Q(x)$. Voici-en un autre.

Trouvons une matrice $Q(x)$ telle que les colonnes de $Z(x)$ varient le plus lentement possible. Une variation régulière de $Z(x)$ permet en général d'intégrer (12) avec un pas plus grand. A cette fin minimisons les normes $\|\mathbf{z}_i'\| = \sqrt{(\mathbf{z}_i', \mathbf{z}_i')}$ en choisissant convenablement les éléments $q_{ij}(x)$ de la matrice $Q(x)$. La colonne $\mathbf{z}_i$ vérifie le système

$$\mathbf{z}_i' = A\mathbf{z}_i - \sum_{j=1}^{r} q_{ij}\mathbf{z}_j.$$

Les vecteurs $\sum_{j=1}^{r} q_{ij}\mathbf{z}_j$ forment un sous-espace vectoriel. Pour que $\|\mathbf{z}_i'\|$ soit minimale, il faut et il suffit que le vecteur $\mathbf{z}_i'$ soit orthogonal à ce sous-espace, i.e. $(\mathbf{z}_i', \mathbf{z}_j) = 0$ pour n'importe quel j. L'ensemble des conditions $(\mathbf{z}_i', \mathbf{z}_j) = 0$, quels que soient i, j, se met sous la forme matricielle

$$Z^TZ' = 0. \tag{16}$$

En y portant la valeur Z' il vient

$$Z^TAZ - Z^TZQ = 0.$$

Si la matrice Z^TZ est inversible, alors $Q = (Z^TZ)^{-1}Z^TAZ$; c'est pourquoi

$$Z' = (E - Z(Z^TZ)^{-1}Z^T)\,AZ. \tag{17}$$

Multiplions (17) par Z^T:

$$Z^TZ' = Z^TAZ - (Z^TZ)(Z^TZ)^{-1}Z^TAZ = 0;$$

la transposition de $Z^TZ' = 0$ donne

$$(Z^T)'Z = 0,$$

d'où

$$(Z^TZ)' = 0,$$

i.e.

$$Z^T(x)\,Z(x) = Z^T(0)\,Z(0). \tag{18}$$

Ainsi, si $Z(x)$ est solution de (17), la matrice $Z^T(x)\,Z(x)$ reste inversible pour tout x à condition de l'être lorsque $x = 0$. Les relations écrites impliquent

l'existence d'une solution de (17) qui vérifie la condition (18) si $Z^T(0)\,Z(0)$ est inversible. Prenons pour colonnes de la matrice $Z(0)$ r solutions linéairement indépendantes du système $Bz = 0$ et trouvons la solution de (17) dans cette condition initiale. $Z^T(0)\,Z(0)$ est la matrice de Gram pour ces r vecteurs linéairement indépendants, elle est donc inversible. La donnée d'un sous-espace W de solutions du système non homogène qui vérifient la condition à l'extrémité gauche exige qu'on se donne également un certain vecteur de ce sous-espace. Si $B\mathbf{u}(0) = \mathbf{b}$, la solution du système $\mathbf{u}' = A\mathbf{u} + \mathbf{f}$ avec la condition initiale $\mathbf{u}(0)$ appartient à W. Tout vecteur $Z(x)\mathbf{q}(x)$, avec $\mathbf{q}(x)$ un vecteur colonne r-dimensionnel, est parallèle à W, c'est pourquoi la solution de

$$\mathbf{u}' = A\mathbf{u} + \mathbf{f} - Z\mathbf{q} \tag{19}$$

pour $\mathbf{q}(x)$ quelconque lui appartient elle aussi. Sous l'hypothèse d'un $\|\mathbf{u}'\|$ minimal sur l'ensemble de tous les $\mathbf{q}$ possibles, nous obtenons un système dont l'écriture matricielle est $Z^T\mathbf{u}' = 0$ ou encore

$$Z^T(A\mathbf{u} + \mathbf{f} - Z\mathbf{q}) = 0.$$

Il en résulte

$$\mathbf{q} = (Z^TZ)^{-1}Z^T(A\mathbf{u} + \mathbf{f}),$$

et le système (19) s'écrit finalement

$$\mathbf{u}' = (E - Z(Z^TZ)^{-1}Z^T)(A\mathbf{u} + \mathbf{f}). \tag{20}$$

La solution cherchée appartient à la variété $\mathbf{u} + Z\mathbf{c}$. La valeur $\mathbf{c}(x)$ se détermine à partir de la condition à l'extrémité droite $D\mathbf{u}(X) = \mathbf{d}$. La substitution de $\mathbf{y} = \mathbf{u} + Z\mathbf{c}$ dans $\mathbf{y}' = A\mathbf{y} + \mathbf{f}$ donne

$$(E - P)(A\mathbf{u} + \mathbf{f}) + Z\mathbf{c}' + (E - P)AZ\mathbf{c} = A\mathbf{u} + AZ\mathbf{c} + \mathbf{f}$$

avec

$$P = Z(Z^TZ)^{-1}Z^T;$$

d'où

$$Z\mathbf{c}' = PAZ\mathbf{c} + PA\mathbf{u} + P\mathbf{f},$$

ou encore

$$\mathbf{c}' = Z^{-1}P(AZ\mathbf{c} + A\mathbf{u} + \mathbf{f}). \tag{21}$$

Résolvons ce système dans le sens des x décroissants, puis déterminons $\mathbf{y}(x) = \mathbf{u}(x) + Z(x)\,\mathbf{c}(x)$.

Laquelle de ces méthodes du balayage différentiel, le procédé (8)-(10) ou celui (17), (20), (21), choisira-t-on de préférence? On ne saurait y répondre catégoriquement sans prendre en considération les propriétés concrètes du système considéré. Dans la seconde méthode les fonctions $Z(x)$, $\mathbf{u}(x)$ correspondantes sont en général par construction plus régulières et, à précision égale du résultat, l'intégration du système (17) s'effectue, paraît-il, avec un pas plus important que celle de (8). Par ailleurs, si l'on intègre le système (17), le volume de calcul par pas peut s'avérer légèrement supérieur à cause de l'inversion nécessaire de la matrice Z^TZ. Il se peut qu'on renonce à inverser celle-ci à chaque pas d'intégration (car cette matrice ne dépend pas de x), surtout si $Z(0)$ a ses colonnes orthonormées et que $Z^T(0)\,Z(0)$ soit donc une matrice unité. Avec de telles simplifications du système de départ (17), (20), (21), ses solutions se conservent; cependant, la sensibilité aux arrondis peut augmenter, ce qui oblige, à moins d'une analyse circonstanciée, à abandonner cette idée.

Dans le cas de l'équation $y'' - p(x)\,y = f(x)$ du § 4, c'est par hasard que la fermeture de l'algorithme s'est avérée régulière et les solutions de ses équations stables par rapport aux arrondis. Il en va autrement pour le système (5.1). La fermeture de l'algorithme de balayage orthogonal (8)-(10) est régulière en ce sens que les solutions des équations (8)-(10) sont toujours bornées par des quantités pas très grandes; par ailleurs, le système d'équations aux variations pour (8) peut posséder des solutions à croissance brusque, ce qui provoque des

doutes quant à la stabilité pratique de l'algorithme. Or, en réalité, l'algorithme est stable vis-à-vis des arrondis. Le fait est que la notion de fermeture d'un algorithme de calcul souffre d'une certaine indétermination. La construction formelle de cette fermeture peut aboutir à diverses équations différentielles relativement à une même fonction et telles que les solutions des unes soient sensibles aux arrondis et les solutions des autres non. La question de la stabilité ne sera tranchée correctement que si la fermeture reflète fidèlement toutes les propriétés de l'algorithme étudié. En ce qui concerne le balayage orthogonal, il serait plus juste d'obtenir, au lieu de (8), l'équation

$$Z' = AZ - Z(Z^TZ)^{-1}R$$

dont les solutions sont réellement insensibles aux arrondis.

Nous ne nous étonnons pas si nos derniers raisonnements dissuadent le lecteur d'étudier les fermetures des algorithmes de calcul. Il ne faut cependant pas oublier que nous proposons de les utiliser lors de l'étude préalable et de la construction de nouveaux algorithmes dans le cas où une analyse exhaustive de l'erreur risque de coûter trop cher en temps et argent. Les propriétés des règles de calcul ainsi obtenues seront ensuite testées par l'expérience numérique.

PROBLÈMES.

1. Construire une méthode numérique de résolution du problème aux limites, dont la fermeture soit l'algorithme (17), (20), (21).

2. Construire les analogues des algorithmes décrits de balayage orthogonal qui soient adaptés aux problèmes aux limites dans le cas du système $\mathbf{y}'' - A(x)\,\mathbf{y} = \mathbf{f}(x)$.

§ 8. Problèmes aux limites non linéaires

Il existe une analogie entre les méthodes de résolution de ces problèmes et celles des problèmes algébriques non linéaires. Comme pour ceux-ci on discutera diverses techniques sans qu'aucune recommandation en sorte pour des problèmes aux limites arbitraires. Tout procédé de résolution d'un problème aux limites non linéaire consiste au fond à résoudre, à sa place, un système d'équations non linéaires. Les procédés diffèrent par le choix des paramètres de ces problèmes auxiliaires et, naturellement, par la manière dont on les aborde.

On venait à bout de nombreux problèmes aux limites tels que le problème de tir sur but ci-dessous bien avant l'avènement des machines à calculer et l'apparition même de la notion d'équation différentielle. Si bien qu'aujourd'hui encore on élabore souvent de nouvelles méthodes par analogie avec les procédés antérieurs à l'ère des ordinateurs.

Soit le cas où la pièce d'artillerie, l'objectif et l'observateur sont dans un même plan horizontal. La trajectoire du projectile est décrite par un système de trois équations différentielles du deuxième ordre avec les conditions aux limites suivantes: l'origine de la trajectoire coïncide avec la pièce et son extrémité avec le point visé. Nous nous abstiendrons d'écrire ce système vu que sa forme concrète est sans importance. Le tireur connaît deux paramètres — l'azimut α du canon et l'angle au niveau β. Si le tir n'a pas été réglé et qu'on ait négligé les conditions atmosphériques, l'obus tiré au premier coup (ce qui correspond à la résolution du problème de Cauchy pour l'approximation d'ordre zéro des conditions initiales cherchées) tombe quelque part à proximité du but. Considérons le cas le plus simple où l'observateur est situé près de la pièce.

En constatant un écart en direction $\Delta\alpha$ entre le but visé et le point de chute réel, il modifie de façon adéquate l'angle α pour qu'au coup suivant les deux directions se confondent. Par suite de nombreux facteurs dont des conditions atmosphériques non constantes, cela n'aura généralement pas lieu encore que toute correction en direction ultérieure s'avère d'ordinaire superflue. Ensuite l'observateur choisit des angles β tels que deux points d'impact soient alignés avec l'objectif de façon à «l'encadrer à une fourchette». Si les appareils nécessaires lui font défaut, la seule chose qu'il puisse dire c'est si le projectile tombe en deçà ou au-delà du but. Si $g(\beta)$ est la portée et g_0 la distance au but, l'observateur résout l'équation $g(\beta) = g_0$ connaissant à chaque coup seulement le signe de la différence $g(\beta) - g_0$. Il choisit alors chaque fois un β qui soit la moyenne arithmétique des deux valeurs β les plus voisines correspondant au coup long ($\operatorname{sg}(g(\beta) - g_0) > 0$) et au coup court ($\operatorname{sg}(g(\beta) - g_0) < 0$).

Le problème se complique singulièrement si le poste d'observation se trouve ailleurs. Le tir de réglage peut s'effectuer alors selon une règle plus compliquée analogue à la méthode de Seidel. L'observateur obtient l'information sur l'écart γ entre l'objectif et le point d'impact tels qu'il les voit du poste et, lorsque $\gamma = 0$, sur la distance relative de l'un et de l'autre. Lorsque l'angle β est fixe, l'observateur corrige l'angle α de sorte que la direction du but et celle du point de chute réel se confondent. L'écart γ est une fonction $\gamma(\alpha, \beta)$. Aussi à cette étape l'observateur résout par itération l'équation $\gamma(\alpha, \beta) = 0$ avec β fixe. Puisque la valeur de $\gamma(\alpha, \beta)$ est connue, on peut appliquer une technique plus rapide que la dichotomie, la méthode de la sécante par exemple. La valeur obtenue de l'angle α dépend de β et nous notons $\alpha(\beta)$ la solution de l'équation $\gamma(\alpha, \beta) = 0$. Si le point d'impact, l'objectif et l'observateur se situent sur une même droite, ce dernier indique le sens de la variation de β; pour cette nouvelle valeur de β on a $\gamma \neq 0$ et il faut préciser α. Si de plus l'objectif est pris en fourchette, l'observateur entreprend de partager en deux l'intervalle contenant la valeur β cherchée. Soient g_0 la distance le séparant du but visé et $g(\alpha, \beta)$ celle qui le sépare du point d'impact pour α et β donnés. Le problème se ramène donc à résoudre le système

$$\gamma(\alpha, \beta) = 0, \quad g(\alpha, \beta) = g_0.$$

Le comportement de l'observateur dans son ensemble se caractérise comme suit. Il résout le problème en choisissant β compte tenu de l'information sur le signe de la différence $g(\alpha(\beta), \beta) - g_0$; ici $\alpha(\beta)$ est solution de $\gamma(\alpha, \beta) = 0$, qui s'obtient à son tour par une méthode itérative. Un observateur expérimenté diminue certes le nombre d'itérations (coups) en combinant intuitivement les méthodes classiques indiquées plus haut.

L'idée de la plupart des procédés de résolution des problèmes aux limites non linéaires est la suivante. Le problème considéré se ramène à résoudre un système non linéaire d'équations par rapport à des inconnues pas très nombreuses; ensuite on applique à ce système une méthode propre à des systèmes non linéaires. Le choix des variables auxiliaires (*paramétrisation du problème*) est souvent décisif.

Soit un problème aux limites non linéaire

$$\begin{gathered} \mathbf{y}' = \mathbf{f}(x, \mathbf{y}), \quad \mathbf{B}(\mathbf{y}(0)) = 0, \quad \mathbf{D}(\mathbf{y}(X)) = 0, \\ \mathbf{y} = (y_1, \ldots, y_l)^T, \quad \mathbf{B} = (b_1, \ldots, b_{l-r})^T, \quad \mathbf{D} = (d_1, \ldots, d_r)^T \end{gathered} \tag{1}$$

et soient les fonctions $g_1(\mathbf{y}(0)), \ldots, g_r(\mathbf{y}(0))$ telles que le système

$$\begin{gathered} b_1(\mathbf{y}(0)) = 0, \ \ldots, \ b_{l-r}(\mathbf{y}(0)) = 0, \\ g_1(\mathbf{y}(0)) = g_1, \ \ldots, \ g_r(\mathbf{y}(0)) = g_r \end{gathered} \tag{2}$$

définit de façon unique [illegible]
Le problème (1) se ramène alors à un système non linéaires relativement aux paramètres $g_1, \ldots, g_r$.

La condition aux limites au point 0 s'écrit assez souvent

$$y_1(0) = 0, \ \ldots, \ y_{l-r}(0) = 0;$$

on prend alors en qualité de $g_1(y(0)), \ldots, g_r(y(0))$ la fonction $y_{l-r+1}(0), \ldots, y_l(0)$. Ce sont donc les composantes inconnues de la solution au point $x = 0$ qui font office de paramètres cherchés. En général, le système nécessaire ne s'explicite pas, mais on connaît l'algorithme qui permet de calculer son second membre.

En résolvant le système (2) par rapport à tous les paramètres $g_1, \ldots, g_r$ on détermine le vecteur $y(0) = \omega_0(g)$; la résolution du problème de Cauchy avec la condition initiale $y(0)$ permet de déterminer $y(X) = \omega_X(g)$, ensuite de trouver $\psi(g) = D(\omega_X(g))$. Ainsi, le problème (1) se réduit à la résolution du système non linéaire

$$\psi(g) = 0. \tag{3}$$

On choisira la méthode de résolution en tenant compte des particularités suivantes de ce système:

1. Le calcul de chaque ψ coûte trop cher — il exige qu'on résolve numériquement le problème de Cauchy; aussi faut-il se tourner vers les procédés dans lesquels ces valeurs sont calculées le moins possible.

2. L'effort de calcul que demande la recherche du vecteur ψ ou celle de l'une de ses composantes ψ_i est pratiquement le même; aussi les méthodes du type de Seidel qui se servent à chaque pas de l'information sur la seule composante ψ_i sont-elles en général inefficaces.

Le tir sur but que nous avons décrit plus haut est analogue à la méthode de Seidel, mais sa légitimité découle du caractère spécifique du problème en raison des possibilités limitées de l'observateur.

Un groupe important de méthodes de résolution des systèmes non linéaires utilisent, en plus des valeurs des fonctions ψ_i, celles de leurs dérivées recherchées de deux manières.

1. Parlons, pour fixer les idées, des dérivées $\frac{\partial \psi_i}{\partial g_1}$ et donnons-nous les valeurs $g_1^0, \ldots, g_r^0$. Trouvons, à l'aide du système (2), les $y_1(0), \ldots, y_l(0)$ correspondants. En dérivant les équations de (2) par rapport à g_1 nous obtenons un système d'équations permettant de trouver $\frac{\partial y_j(0)}{\partial g_1}$:

$$\sum_{j=1}^{l} \frac{\partial b_k}{\partial y_j(0)} \frac{\partial y_j(0)}{\partial g_1} = 0, \quad k = 1, \ \ldots, \ l-r,$$

$$\sum_{j=1}^{l} \frac{\partial g_m}{\partial y_j(0)} \frac{\partial y_j(0)}{\partial g_1} = \delta_{m-1}, \quad m = 1, \ \ldots, \ r.$$

de dérivées

$$\frac{\partial y(x)}{\partial g_1} = \left(\frac{\partial y_1(x)}{\partial g_1}, \ldots, \frac{\partial}{\partial g_1}\right)$$

ifie le système d'équations différentielles

$$\frac{d}{dx}\left(\frac{\partial y(x)}{\partial g_1}\right) = f_y(x, y)\frac{\partial y(x)}{\partial g_1}, \quad f_y = \left[\frac{\partial f_i}{\partial y_j}\right]. \tag{4}$$

En le résolvant numériquement, on obtient le vecteur $\frac{\partial y(X)}{\partial g_1}$. Maintenant, on peut calculer

$$\frac{\partial \psi_i}{\partial g_1} = \sum_{j=1}^{l} \frac{\partial d_i}{\partial y_j(X)} \frac{\partial y_j(X)}{\partial g_1} = \left(\operatorname{grad} d_i, \frac{\partial y(X)}{\partial g_1}\right). \tag{5}$$

Ce mode opératoire s'avère avantageux si, par exemple, on résout à plusieurs reprises le problème (1) avec un même second membre $f(x, y)$ et dans des conditions aux limites différentes.

Par contre, en tant que méthode usuelle de définition des dérivées $\frac{\partial \psi_i}{\partial g_j}$, il présente le défaut suivant: l'utilisateur doit rédiger, en plus du programme de calcul des seconds membres de l'équation, le programme de calcul des dérivées $\frac{\partial f_i}{\partial y_j}$.

2. Dans la plupart des méthodes résolvant des systèmes d'équations par itération, les valeurs des dérivées des seconds membres doivent être obtenues avec une précision modérée. On les trouve donc par des formules de dérivation numérique, par exemple par la formule simple

$$\frac{\partial \psi_i(g_1, \ldots, g_r)}{\partial g_1} \approx \frac{\psi_i(g_1 + \Delta, g_2, \ldots, g_r) - \psi_i(g_1, \ldots, g_r)}{\Delta}.$$

Le calcul des $\frac{\partial \psi_i}{\partial g_1}$ moyennant cette formule exige (par rapport à la recherche des valeurs $\psi_i(g_1, \ldots, g_r)$) une résolution numérique supplémentaire du problème de Cauchy (correspondant aux paramètres $g_1 + \Delta, g_2, \ldots, g_r$). Comme le calcul des valeurs des fonctions ψ_i est susceptible de donner lieu à une erreur notable qui ne fait qu'augmenter avec la division par Δ, un choix rationnel de Δ ne se fait souvent qu'au bout d'une analyse complémentaire. La méthode décrite se caractérise par le calcul simultané des dérivées par rapport à la variable g_j fixe de toutes les fonctions ψ_i ; il paraît que dans ce cas la tactique raisonnable consiste à mettre en œuvre des processus itératifs dans lesquels les paramètres g_j sont améliorés à tour

de rôle. On pourra imaginer le processus suivant: la précision des paramètres g_j s'améliore de façon cyclique, on déduit une valeur plus exacte $\bar{g}_j$ à partir de la condition de minimum d'une certaine fonction

$$\Phi_j(\psi_1(g_1, \ldots, g_r) + \frac{\partial \psi_1(g_1, \ldots, g_r)}{\partial g_j}(\bar{g}_j - g_j), \ldots, \psi_r(g_1, \ldots, g_r) + \frac{\partial \psi_r(g_1, \ldots, g_r)}{\partial g_j}(\bar{g}_j - g_j)$$

(voir pour plus de détails § 4, chap. VII).

Comme dans le cas des problèmes aux limites linéaires, on s'interroge sur l'applicabilité de la méthode en présence d'arrondis réels. Lorsque les solutions η du système d'équations aux variations $\eta' = f_y\eta$ croissent fortement avec x, les erreurs dont sont affectés g_j et les erreurs de calcul dans l'intégration numérique engendrent de grosses erreurs sur les valeurs de la fonction ψ_i, ce qui entraîne en fin de compte une imprécision importante de la solution obtenue. Si cette dernière erreur s'avère inadmissible, on choisit d'autres variables auxiliaires. Décrivons des procédés de résolution dans lesquels les paramètres formels sont les valeurs de la solution sur tout le segment.

Nous avons déjà proposé le mode opératoire suivant pour le problème non linéaire $P(u) = 0$: nous avons choisi un opérateur linéaire Lu agissant dans le même espace que P. Les approximations suivantes étaient déterminées par l'égalité

$$L(u_{n+1} - u_n) + P(u_n) = 0.$$

Ainsi, en ce qui concerne le problème (1), on peut prendre un opérateur L tel que

$$\begin{aligned} L\mathbf{y} &= C(x)\mathbf{y}' - A(x)\mathbf{y}, \quad 0 < x < X, \\ L\mathbf{y}|_{x=0} &= B_0\mathbf{y}(0), \\ L\mathbf{y}|_{x=X} &= D_0\mathbf{y}(X), \end{aligned}$$

$C(x)$, $A(x)$, B_0, D_0 étant des matrices. Les approximations suivantes s'obtiennent alors comme les solutions du système linéaire

$$\begin{aligned} C(x)(\mathbf{y}_{n+1} - \mathbf{y}_n)' - A(x)(\mathbf{y}_{n+1} - \mathbf{y}_n) + \mathbf{y}_n' - \mathbf{f}(x, \mathbf{y}_n) &= 0, \\ B_0(\mathbf{y}_{n+1}(0) - \mathbf{y}_n(0)) + \mathbf{B}(\mathbf{y}_n(0)) &= 0, \\ D_0(\mathbf{y}_{n+1}(X) - \mathbf{y}_n(X)) + \mathbf{D}(\mathbf{y}_n(X)) &= 0. \end{aligned} \tag{6}$$

Dans le cas de la méthode de Newton (cf. § 2, chap. VII), L dépend de n et est un opérateur de dérivée correspondant à l'approximation $\mathbf{y}_n(x)$:

$$\begin{aligned} L\mathbf{u} &= \mathbf{u}' - \mathbf{f}_y(x, \mathbf{y}_n(x))\mathbf{u} \quad \text{pour} \quad x \in (0, X), \\ L\mathbf{u}|_{x=0} &= B_y(\mathbf{y}_n(0))\mathbf{u}(0), \\ L\mathbf{u}|_{x=X} &= D_y(\mathbf{y}_n(X))\mathbf{u}(X), \end{aligned}$$

où

$$B_y = \left[\frac{\partial b_i}{\partial y_j(0)}\right], \quad D_y = \left[\frac{\partial d_i}{\partial y_j(X)}\right].$$

Les relations (6) s'écrivent alors

$$\begin{aligned} &y'_{n+1} - f_y(x, y_n(x))(y_{n+1} - y_n) - f(x, y_n(x)) = 0, \quad 0 < x < X, \\ &B_y(y_n(0))(y_{n+1}(0) - y_n(0)) + B(y_n(0)) = 0, \\ &D_y(y_n(0))(y_{n+1}(0) - y_n(0)) + D(y_n(0)) = 0. \end{aligned} \tag{7}$$

D'après les théorèmes généraux sur la convergence de la méthode de Newton, le processus itératif (5) converge pour $f(x, y)$, b_i, d_i suffisamment réguliers si l'approximation initiale est suffisamment proche de la solution cherchée. Il est naturel que les méthodes de résolution des problèmes linéaires auxiliaires, du problème (7) par exemple, doivent être assez stables par rapport à l'erreur de calcul.

Nous venons de décrire un procédé où le rôle des paramètres cherchés est formellement joué par les valeurs de la solution sur tout le segment d'intégration. Comme le problème aux limites linéaire se résout numériquement, les calculs réels porteront sur les approximations y_n^k des valeurs de la solution $y(x_k)$ aux nœuds $0 = x_0 < x_1 < \ldots < x_N = X$ du réseau. Ainsi, en intégrant par la méthode d'Euler, on obtient, au lieu de (7):

$$\begin{aligned} &\frac{y_{n+1}^{j+1} - y_{n+1}^j}{x_{j+1} - x_j} - f_y(x_j, y_n^j)(y_{n+1}^j - y_n^j) - f(x_j, y_n^j) = 0, \quad 0 \leqslant j < N, \\ &B_y(y_n^0)(y_{n+1}^0 - y_n^0) + B(y_n^0) = 0, \\ &D_y(y_n^N)(y_{n+1}^N - y_n^N) + D(y_n^N) = 0. \end{aligned}$$

Cette paramétrisation du problème aux limites a l'inconvénient d'encombrer la mémoire si le nombre de points d'intégration est important et le système d'ordre élevé. Il se peut que les approximations $y_n(x)$ offrent une régularité différente (la régularité de certaines approximations nécessite un pas d'intégration faible sur tels secteurs tandis que la régularité d'autres approximations sur d'autres secteurs). La méthode est alors modernisée comme l'exige le passage d'un réseau sur un autre.

Lorsque nous avons eu à résoudre des systèmes d'équations non linéaires, nous avons quelquefois recommandé de réduire le volume de calcul en itérant par la méthode de Newton pour des valeurs inexactes des dérivées dans les seconds membres. En agissant de façon analogue dans le cas du problème (1) nous diminuons l'occupation de mémoire rapide. Donnons-nous des points

$$0 = X_0 < X_1 < \ldots < X_m = X$$

et mettons en réserve, au bout de chaque pas d'itération, les seules approximations Y_j^n, $j = 0, \ldots, m-1$, de la solution en $X_0, \ldots$

. . ., X_{m-1}. Si les solutions du système d'équations aux variations sur la solution cherchée ne croissent pas considérablement sur les segments $[X_{j-1}, X_j]$, une fois tous les $\mathbf{y}(X_j)$ trouvés, les valeurs de la solution sur tout le segment s'obtiennent par intégration numérique du système de départ sur les segments $[X_{j-1}, X_j]$.

Décrivons cette technique plus en détail. Soient $Y = (\mathbf{y}_0, \ldots$ $\ldots, \mathbf{y}_{m-1})$, P un opérateur faisant correspondre à toute fonction $z(x)$ la matrice $Pz = (z(X_0), \ldots, z(X_{m-1}))$ et Q un opérateur qui fait correspondre à la matrice Y une fonction $\mathbf{u}(x) = QY$ définie sur $[0, X]$, de plus $QP\mathbf{y} = \mathbf{y}$ si $\mathbf{y}$ est solution du problème initial. Supposons qu'il existe un processus itératif de départ tel que l'approximation suivante se déduise de la précédente comme solution du système

$$\mathbf{Y}'_{n+1} = \mathbf{A}(x, \mathbf{Y}_n, \mathbf{Y}_{n+1})$$

avec les conditions aux limites $\mathbf{b}(\mathbf{Y}_n(0), \mathbf{Y}_{n+1}(0)) = 0$, $\mathbf{d}(\mathbf{Y}_n(X), \mathbf{Y}_{n+1}(X)) = 0$, la solution cherchée étant également un point stationnaire du processus par itérations:

$$\mathbf{y}' = \mathbf{A}(x, \mathbf{y}, \mathbf{y}),$$
$$b(\mathbf{y}(0), \mathbf{y}(0)) = 0, \quad d(\mathbf{y}(X), \mathbf{y}(X)) = 0.$$

Cette solution est évidemment un point stationnaire pour

$$\begin{aligned} \mathbf{Y}'_{n+1} &= \mathbf{A}(x, QP\mathbf{Y}_n, \mathbf{Y}_{n+1}), \\ \mathbf{b}(QP\mathbf{Y}_n|_{x=0}, \mathbf{Y}_{n+1}(0)) &= 0, \\ \mathbf{d}(QP\mathbf{Y}_n|_{x=X}, \mathbf{Y}_{n+1}(X)) &= 0, \end{aligned} \tag{8}$$

i.e.

$$\mathbf{y}' = \mathbf{A}(x, QP\mathbf{y}, \mathbf{y}), \quad \mathbf{b}(QP\mathbf{y}|_{x=0}, \mathbf{y}) = 0, \quad \mathbf{d}(QP\mathbf{y}|_{x=X}, \mathbf{y}) = 0.$$

Voici une autre mise en œuvre de ce processus itératif. On prend pour Q un opérateur faisant correspondre à la matrice $Y =$ $= (\mathbf{Y}_0, \ldots, \mathbf{Y}_{m-1})$ une fonction $\mathbf{u}(x)$ qui sur les intervalles semi-fermés $[X_0, X_1), \ldots, [X_{m-2}, X_{m-1})$ et sur le segment $[X_{m-1}, X_m]$ est solution du problème de Cauchy pour le système $\mathbf{u}' = f(x, \mathbf{u})$ dans les conditions initiales $\mathbf{u}(X_j) = \mathbf{Y}_j$. Supposons que

$$\mathbf{A}(x, \mathbf{y}_n, \mathbf{y}_{n+1}) = \mathbf{g}_0(x, \mathbf{y}_n) + \mathbf{g}_1(x, \mathbf{y}_n)\,\mathbf{y}_{n+1},$$
$$\mathbf{b}(\mathbf{y}_n(0), \mathbf{y}_{n+1}(0)) = \mathbf{b}_0(\mathbf{y}_n(0)) + \mathbf{b}_1(\mathbf{y}_n(0))\,\mathbf{y}_{n+1}(0),$$
$$\mathbf{d}(\mathbf{y}_n(X), \mathbf{y}_{n+1}(X)) = \mathbf{d}_0(\mathbf{y}_n(X)) + \mathbf{d}_1(\mathbf{y}_n(X))\,\mathbf{y}_{n+1}(X).$$

Le problème

$$\begin{aligned} \mathbf{y}'_{n+1} &= \mathbf{g}_0(x, QP\mathbf{y}_n) + \mathbf{g}_1(x, QP\mathbf{y}_n)\,\mathbf{y}_{n+1}, \\ \mathbf{b}_0((QP\mathbf{y}_n)|_{x=0}) + \mathbf{b}_1((QP\mathbf{y}_n)|_{x=0})\,\mathbf{y}_{n+1}(0) &= 0, \\ \mathbf{d}_0((QP\mathbf{y}_n)|_{x=X}) + \mathbf{d}_1((QP\mathbf{y}_n)|_{x=X})\,\mathbf{y}_{n+1}(X) &= 0 \end{aligned} \tag{9}$$

se résout alors par une méthode valable pour les problèmes aux limites linéaires.

Appliquons la première variante du balayage orthogonal à condition que l'orthogonalisation de la solution s'effectue justement aux points $X_1, \ldots, X_{m-1}$. On demande alors de garder en mémoire les valeurs de certains vecteurs et matrices correspondant aux points X_j et d'intégrer numériquement sur les segments $[X_j, X_{j+1}]$ plusieurs solutions du système homogène

$$\mathbf{v}' = \mathbf{g}_1(x, QP\mathbf{y}_n)\,\mathbf{v} \tag{10}$$

et une solution du système non homogène (9). Puisque $QP\mathbf{y}_n$ est par définition solution du système

$$(QP\mathbf{y}_n)' = \mathbf{f}(x, QP\mathbf{y}_n), \tag{11}$$

(9)-(11) sont intégrés tous à la fois sans mettre en réserve les valeurs de ces fonctions en des points intermédiaires.

Le présent algorithme est susceptible d'un raffinement qui présente un intérêt certain dans la résolution d'autres problèmes linéaires et non linéaires.

Les valeurs approchées de la solution s'obtiennent sous forme de combinaisons linéaires de solutions particulières des systèmes (9) et (10). S'il y a convergence vers la solution, $QP\mathbf{y}_n(x) \to \mathbf{y}(x)$ et les systèmes (9), (10) deviennent

$$\mathbf{u}' = \mathbf{g}_0(x, \mathbf{y}) + \mathbf{g}_1(x, \mathbf{y})\,\mathbf{u},$$
$$\mathbf{v}' = \mathbf{g}_1(x, \mathbf{y})\mathbf{v}.$$

En appliquant la méthode de Newton, ils s'écrivent par exemple

$$\mathbf{u}' = \mathbf{f}(x, \mathbf{y}(x)) + f_y(x, \mathbf{y}(x))\,(\mathbf{u} - \mathbf{y}(x)),$$
$$\mathbf{v}' = f_y(x, \mathbf{y}(x))\,\mathbf{v}.$$

La solution cherchée peut être (c'est le cas, par exemple, des problèmes avec couche limite) assez régulière sauf au voisinage des extrémités du segment d'intégration; en même temps, les solutions du système (10) varient brusquement sur tout le segment. Si bien que même là où la solution cherchée est régulière, pour l'obtenir avec une précision acceptable on doit intégrer les systèmes auxiliaires (10), (11) avec un pas très petit. On l'évite en passant à une nouvelle fonction inconnue

$$\mathbf{r}_{n+1} = \mathbf{y}_{n+1} - QP\mathbf{y}_n.$$

Pour fixer les idées, on raisonnera sur la méthode de Newton. Aux points où $\mathbf{r}_{n+1}(x)$ est continûment dérivable on a

$$\mathbf{r}'_{n+1} = \mathbf{f}(x, QP\mathbf{y}_n) - (QP\mathbf{y}_n)' + f_y(x, QP\mathbf{y}_n)\,\mathbf{r}_{n+1}$$

ou encore

$$\mathbf{r}'_{n+1} = f_y(x, QP\mathbf{y}_n)\,\mathbf{r}_{n+1}. \tag{12}$$

Les conditions aux limites correspondantes sont

$$\begin{aligned} &B((QP\mathbf{y}_n)|_{x=0}) + B_y((QP\mathbf{y}_n)|_{x=0})\,\mathbf{r}_{n+1}(0) = 0,\\ &D((QP\mathbf{y}_n)|_{x=X}) + D_y((QP\mathbf{y}_n)|_{x=X})\,\mathbf{r}_{n+1}(X) = 0. \end{aligned} \tag{13}$$

La fonction $\mathbf{y}_{n+1}$ est définie comme une solution continûment dérivable de (9) et $QP\mathbf{y}_n$ présente peut-être des discontinuités en $X_1, \ldots, X_{m-1}$. Aussi

$$[\mathbf{r}_{n+1}]_{X_j} = -[QP\mathbf{y}_n]_{X_j}, \quad j = 1, \ldots, m-1;$$

ici

$$[g]_x = \lim_{\varepsilon \to +0} (g(x+\varepsilon) - g(x-\varepsilon)).$$

Le système d'équations donnant $r_{n+1}(x)$ s'écrit sous forme généralisée

$$r'_{n+1} = g_n(x) + f_y(x, QPy_n)\, r_{n+1}, \tag{14}$$

$$g_n(x) = -\sum_{j=1}^{m-1} [QPy_n]_{X_j}\, \delta(x - X_j);$$

avec $\delta(t)$ la fonction de Dirac. En résolvant le problème aux limites (13)-(14) par la méthode du balayage orthogonal à points d'orthogonalisation $X_1, \ldots$ $\ldots, X_{m-1}$, la présence des fonctions de Dirac dans les seconds membres du système ne crée aucune complication: après avoir intégré l'équation homogène (10) sur les segments $[X_{j-1}, X_j]$, on ajoute $-[QPy_n]_{X_j}$ à la valeur obtenue.

Le système homogène associé à (9) et (14) étant le même, on dirait qu'une telle substitution de la fonction inconnue ne change rien. En réalité, ces deux processus itératifs diffèrent de façon essentielle. Dans le premier cas, on trouve les approximations $y_{n+1}(x)$ sur chaque $[X_{j-1}, X_j]$ sous forme de combinaisons linéaires $u_n(x) + \sum_{i=1}^{r} c_{in} v_{in}(x)$ des solutions des équations (9) et (10), $u_n(x)$ et, corrélativement, le vecteur de coefficients $c_{in} \not\to 0$ avec le second membre de (9). Dans le deuxième cas

$$r_{n+1} = u_n(x) + \sum_{i=1}^{r} c_{in} v_{in}(x).$$

Les vecteurs $u_n(X_j)$, $v_{in}(X_j)$ étant orthogonaux et $r_{n+1}(X_j) \to 0$, on a $u_n(x) \to 0$ et $c_{in} \to 0$ lorsque n tend vers l'infini. Les fonctions $v_{in}(x)$ sont multipliées par des quantités petites, c'est pourquoi les erreurs commises lors de leur recherche numérique influent bien moins sur les valeurs approchées obtenues. De même, s'il s'agit d'une équation non homogène en $u_n(x)$, on s'attend à des contraintes moins sévères sur le pas d'intégration vu que $u_n(x)$ tend vers zéro.

Dans le cas d'une intégration « grossière » des équations pour les fonctions u_n, v_{in}, le processus itératif subit certaines modifications, et il se peut que la vitesse de convergence se ralentisse. On a cependant des raisons d'espérer que les qualités de l'algorithme s'amélioreront dans leur ensemble par suite d'un volume de calcul plus faible par pas d'itération. Comme les solutions particulières des systèmes (12), (14) sont définies pas très exactement, le système (12) admet, à part l'intégration numérique directe, des procédés plus grossiers. Le calcul des solutions des systèmes homogène (12) et non homogène (14) aux points X_j avec des conditions données en X_{j-1} exige qu'on intègre numériquement les équations aux variations $\eta' = f_y(x, QPy_n)\,\eta$. QPy_n vérifie par définition sur $[X_{j-1}, X_j]$ le système initial

$$(QPy_n)' = f(x, QPy_n).$$

Soit $y_\varepsilon(x)$ une solution de celui-ci dans la condition initiale

$$y_\varepsilon(X_{j-1}) = QPy_n\,|_{X_{j-1}} + \varepsilon v_{j-1}.$$

En établissant un système d'équations aux variations on a vu que la fonction vectorielle $(y_\varepsilon(x) - QPy_n)\,\varepsilon^{-1}$, ε faible, approche la solution $R(x)$ de l'équation (12) sous la condition initiale $R(X_{j-1}) = v_{j-1}$. Sur chacun des segments $[X_{j-1}, X_j]$ on demande de trouver les $r+1$ solutions du système $R' = f_y(x, QPy_n)\,R$. On le fait par le procédé décrit en intégrant le système $y' = f(x, y)$ $r+2$ fois.

Nos raisonnements (chap. VII) sur l'avantage que présente l'amélioration de la précision à mesure qu'on approche la solution, restent valables pour les méthodes itératives. Les développements relatifs à la dernière technique décrite nous montrent cependant qu'on n'augmentera sensiblement la précision qu'en calculant les seconds membres des équations non homogènes, i.e. les quantités $[QPy]\,|_{X_j}$; il suffit donc d'élever la précision en intégrant

$$(QPy_n)' = f(x, QPy_n).$$

Comme pour les problèmes algébriques non linéaires, on pourrait citer de nombreux procédés qui convergent vers la solution à condition d'avoir au départ une approximation assez bonne; cependant, on ne saurait proposer de méthode efficace pour tout problème aux limites non linéaire. Aussi en abordant un problème concret on procède à une « mise au point » de la technique de résolution: on choisit une méthode spéciale donnant l'approximation initiale, on la modernise afin d'élargir le domaine de conditions initiales dans lesquelles elle converge pour classe concrète de problèmes, etc.

§ 9. Approximations du type spécial

Soit le problème aux limites

$$(k(x)\,y'_x)'_x - p(x)\,y = f(x), \quad y(0) = a,\ y(X) = b, \qquad (1)$$

avec $p(x) \geqslant 0$, $k(x)$ trois fois et $p(x)$, $f(x)$ deux fois continûment dérivables sauf en un nombre fini de points en lesquels ces fonctions ou leurs dérivées k', k'', k''', p', p'', f', f'' peuvent posséder des discontinuités de première espèce.

Soient K, P, F les ensembles de points de discontinuité des fonctions k, p, f ou de leurs dérivées et $\Omega = K \cup P \cup F$.

Appelons solution du problème (1) une fonction $y(x)$ telle

1) qu'elle soit continue sur $[0, X]$;

2) qu'elle vérifie l'équation partout sur $[0, X]$, sauf peut-être aux points de l'ensemble Ω;

3) que la fonction $w(x) = k(x)\,y'_x$ soit continue sur $[0, X]$.

Cette dernière condition entraîne que $y'_x = \frac{w(x)}{k(x)}$ est continue partout, à l'exception des points de discontinuité de la fonction $k(x)$ en lesquels elle présente des discontinuités de première espèce.

On montre que les dérivées $y'''(x)$ et $y^{(4)}(x)$ sont continues et uniformément bornées sur l'ensemble $[0, X] \setminus \Omega$.

Si la construction du schéma aux différences néglige le caractère discontinu de y'_x, on risque d'aboutir à un problème aux différences dont la solution ne converge pas vers celle de (1).

Si l'on a à construire des schémas aux différences finies dans le cas d'équations dont les coefficients et les solutions présentent des discontinuités ou d'autres singularités, il est utile d'introduire de

nouvelles variables (fonctions ou arguments) de façon que la solution du nouveau problème soit plus régulière. Comme nous allons le voir, ce passage n'est pas toujours réalisé de façon explicite.

Puisque $w(x)$ est continue et la dérivée

$$w'_x = p(x)\, y + f(x) \tag{2}$$

l'est également, sauf aux points de discontinuité des fonctions $p(x)$ et $f(x)$, la fonction $w(x)$ est continûment dérivable. Notons $x_s = sh$ pour s entier ou non. L'intégration de (2) entre $x_{q-1/2}$ et $x_{q+1/2}$ donne

$$w(x_{q+1/2}) - w(x_{q-1/2}) = \int_{x_{q-1/2}}^{x_{q+1/2}} (p(x)\, y(x) + f(x))\, dx\,; \tag{3}$$

$y(x)$ étant dérivable par morceaux, il vient

$$y(x) = y(x_q) + O(h) \text{ lorsque } x - x_q = O(h).$$

L'égalité (3) se récrit donc

$$w(x_{q+1/2}) - w(x_{q-1/2}) = \int_{x_{q-1/2}}^{x_{q+1/2}} (p(x)\, y(x_q) + f(x) + O(h))\, dx =$$

$$= y(x_q)\, \bar{p}_q + \bar{f}_q + O(h^2),$$

où

$$\bar{p}_q = \frac{1}{h} \int_{x_{q-1/2}}^{x_{q+1/2}} p(x)\, dx, \quad \bar{f}_q = \frac{1}{h} \int_{x_{q-1/2}}^{x_{q+1/2}} f(x)\, dx.$$

Divisons par h ; nous trouvons

$$\frac{w(x_{q+1/2}) - w(x_{q-1/2})}{h} - \bar{p}_q y(x_q) = \bar{f}_q + O(h). \tag{4}$$

Si l'intervalle (x_{j-1}, x_j) renferme des points de discontinuité de y'_x, la substitution directe de $k(x_{j-1/2}) \frac{y(x_j) - y(x_{j-1})}{h^2}$ à $\frac{w(x_{j-1/2})}{h}$ peut donner lieu à une erreur en h^{-1}.

On améliore l'approximation en faisant intervenir la variable indépendante auxiliaire $t = \int_0^x \frac{dx}{k(x)}$. Le caractère borné de la dérivée par rapport à x et l'égalité $\frac{du}{dt} = k(x) \frac{du}{dx}$ entraînent que la dérivée par rapport à t est bornée. La fonction $w = k(x) \frac{dy}{dx} = \frac{dy}{dt}$ a sa dérivée par rapport à x bornée, il en est donc de même de sa dérivée par

rapport à t. Ainsi, la dérivée seconde $y_{tt}^{\sim}$ est bornée et

$$k(x)\frac{dy}{dx}\Big|_{x_{j-1/2}} = \frac{dy}{dt}\Big|_{t_{j-1/2}} = \frac{y(t_j)-y(t_{j-1})}{t_j-t_{j-1}} + O(t_j-t_{j-1}),$$

avec

$$t_n = \int_0^{x_n} \frac{dx}{k(x)}; \quad t_j - t_{j-1} = \int_{x_{j-1}}^{x_j} \frac{dx}{k(x)} = O(h).$$

C'est pourquoi

$$w(x_{j-1/2}) = k(x)\frac{dy}{dx}\Big|_{x_{j-1/2}} = \frac{y(x_j)-y(x_{j-1})}{t_j-t_{j-1}} + O(h). \tag{5}$$

Portons $k(x)\frac{dy}{dx}\Big|_{x_q \pm 1/2}$ à gauche dans (4), il vient

$$r_q = \frac{\dfrac{y(x_{q+1})-y(x_q)}{h\bar{k}_{q+1/2}^{-1}} - \dfrac{y(x_q)-y(x_{q-1})}{h\bar{k}_{q-1/2}^{-1}}}{h} - \bar{p}_q y(x_q) - \bar{f}_q = O(1), \tag{6}$$

où

$$\bar{k}_{j+1/2}^{-1} = \frac{1}{h}\int_{x_j}^{x_{j+1}} \frac{dx}{k(x)};$$

le schéma aux différences finies correspondant a la forme

$$L(y_q) = \frac{\dfrac{y_{q+1}-y_q}{h\bar{k}_{q+1/2}^{-1}} - \dfrac{y_q-y_{q-1}}{h\bar{k}_{q-1/2}^{-1}}}{h} - \bar{p}_q y_q = \bar{f}_q. \tag{7}$$

Son erreur d'approximation est de $O(1)$ au maximum et on l'évalue donc moyennant des considérations supplémentaires.

Dans l'hypothèse de l'absence d'arrondis, l'erreur sur la solution approchée s'écrit, conformément à (2.10):

$$R_n = -h\sum_{k=1}^{N-1} G_n^k r_k.$$

La définition de G_n^k (cf. § 2) et les estimations du même paragraphe entraînent

$$\begin{aligned} &\max_{0\leqslant kh,\ nh\leqslant X} |G_n^k| \leqslant D < \infty, \\ &\max_{0\leqslant kh,\ nh\leqslant X} |G_n^{k+1} - G_n^k| = O(h). \end{aligned} \tag{8}$$

La théorie des équations différentielles dit que les dérivées de la solution jusqu'à l'ordre 4 inclusivement sont uniformément bornées partout, sauf peut-être aux points de Ω. Si l'intervalle

(x_{q-1}, x_{q+1}) ne contient donc pas de points de cet ensemble, l'estimation (6) se prête à une amélioration: dans le premier membre de (6) on développe $y(x)$, $k(x)$, $p(x)$ et $f(x)$ au point x_q par la formule de Taylor avec des restes en h^4, h^3, h^2 et h^2 respectivement. Des calculs élémentaires tenant compte de l'égalité (1) fournissent $r_q = O(h^2)$.

Cette estimation permet d'écrire

$$R_n = O(h^2) - \sum_k{}' h G_n^k r_k.$$

Le signe $\sum_k{}'$ signifie que la sommation s'effectue sur les seuls k tels que l'intervalle (x_{k-1}, x_{k+1}) contient les points de discontinuité des coefficients de l'équation ou de leurs dérivées d'ordre peu élevé correspondant. Pour les valeurs $q = k$ correspondant à $\sum'$ mettons r_q sous la forme

$$r_q = \frac{\dfrac{y(x_{q+1}) - y(x_q)}{h\bar{k}_{q+1/2}^{-1}} - \dfrac{y(x_q) - y(x_{q-1})}{h\bar{k}_{q-1/2}^{-1}}}{h} - \bar{p}_q y(x_q) - \bar{f}_q =$$

$$= \frac{\omega_{q+1/2} - \omega_{q-1/2}}{h} + \left\{ \frac{w(x_{q+1/2}) - w(x_{q-1/2})}{h} - \bar{p}_q y(x_q) - \bar{f}_q \right\},$$

où

$$\omega_{j+1/2} = \frac{y(x_{j+1}) - y(x_j)}{h\bar{k}_{j+1/2}^{-1}} - w(x_{j+1/2}).$$

Selon (4), l'expression entre accolades est égale à $O(h)$, les points de discontinuité sont en nombre fini, et l'erreur s'écrit donc

$$R_n = O(h^2) - \sum_k{}' G_n^k (\omega_{k+1/2} - \omega_{k-1/2}).$$

Si l'intervalle (x_k, x_{k+1}) renferme un point de discontinuité, la valeur $\omega_{k+1/2}$ correspondante sera précédée sous le signe $\sum'$ du coefficient $G_n^k - G_n^{k+1}$; dans le cas contraire, on s'assure par un développement taylorien direct que cette valeur vaut $O(h^2)$. L'expression de R_n est donc susceptible d'être transformée:

$$R_n = O(h^2) + \sum_k{}'' (G_n^{k+1} - G_n^k)\, \omega_{k+1/2},$$

ici $\sum''$ désigne la sommation sur des k correspondant aux intervalles (x_k, x_{k+1}) renfermant des points de discontinuité de $k(x)$. Conformément à (5) $\omega_{k-1/2} = O(h)$ uniformément en tous les k. Recourons également à (8) et obtenons une estimation définitive de l'erreur:

$$\max_{0 \leqslant nh \leqslant X} |R_n| = O(h^2)$$

Dans le cas du schéma (7) le calcul est indépendant de la position des points de discontinuité; il s'agit donc d'un *schéma homogène*.

Son défaut apparent est que ses coefficients $\overline{k}_{q+1/2}$, $\overline{p}_q$, $\overline{f}_q$ s'écrivent comme des intégrales. En réalité, on montre que si l'erreur sur les valeurs de ces coefficients est $O(h^2)$, il en est de même de l'erreur sur la solution approchée. Aussi, dans le cas d'un intervalle (x_{q-1}, x_{q+1}) ne contenant pas de points de discontinuité, on remplace $\overline{p}_q$ par $p(x_q)$, $\overline{f}_q$ par $f(x_q)$ et $\overline{k}_{q\pm1/2}$ par $k(x_{q\pm1/2})$ sans pour autant diminuer la précision.

Nous avons écrit l'erreur d'approximation des méthodes des différences finies de l'intégration numérique sous la forme $E_m y^{(m+1)}(\zeta)$; elle était donc égale à zéro dans le cas où la solution était un polynôme de degré m. Il se peut que la solution du problème soit mieux approximée non pas par des polynômes mais par des combinaisons linéaires de certaines fonctions autres que les monômes x^i. On sait par exemple que les solutions de l'équation différentielle de Bessel $x^2y'' + xy' + (x^2 - \nu^2)\, y = 0$ s'écrivent pour des x grands:

$$(a_0x^{-1/2} + a_1x^{-3/2} + \ldots)\cos x + (b_0x^{-1/2} + b_1x^{-3/2} + \ldots)\sin x.$$

Supposons que la résolution exige la mise en réserve des tableaux des fonctions de Bessel. Parfois on les forme au mieux par intégration numérique de l'équation de Bessel en se servant, pour x important, des approximations aux différences exactes le long des fonctions

$$x^{-k-1/2}\cos x, \quad x^{-k-1/2}\sin x, \quad k = 0, \ldots, l.$$

Un autre procédé commode consiste à approcher, dans le voisinage de chaque point, l'équation considérée par une équation différentielle intégrable sous forme explicite et à construire un schéma aux différences exact pour ses solutions. Pages 491-492 nous avons proposé un algorithme de ce type.

Traitons le cas de l'équation différentielle

$$\mu^2y'' + p(x)\, y = f(x),$$

avec μ petit et $p > 0$.

Lorsque $p = \text{const}$, les solutions $\exp\left(\pm i\frac{x\sqrt{p}}{\mu}\right)$ de cette équation subissent des oscillations de période $2\pi\mu/\sqrt{p}$, c'est-à-dire très grandes. La mesure caractéristique de la variation d'une solution (i. e. l'intervalle sur lequel elle varie de façon substantielle) est de l'ordre de $\mu/\sqrt{p}$, de sorte que si l'on ne fait pas intervenir la spécificité de l'équation donnée et que l'on veuille obtenir une précision élevée, il faut que soit réalisée la condition assez encombrante $h \ll \mu/\sqrt{p}$.

Dans le voisinage de chaque point x_n l'équation examinée est voisine de $\mu^2y'' + p_ny = f_n$, $p_n = p(x_n)$, $f_n = f(x_n)$. La solution

générale de celle-ci s'écrit

$$y(x)=D_1\exp\left(i\frac{(x-x_n)\sqrt{p_n}}{\mu}\right)+D_2\exp\left(-i\frac{(x-x_n)\sqrt{p_n}}{\mu}\right)+\frac{f_n}{p_n}; \quad (9)$$

trouvons un schéma du type

$$a_n y_{n+1}+b_n y_n+c_n y_{n-1}-d_n=0 \quad (10)$$

exact le long de toutes les solutions pareilles. A cette fin, portons (9) dans (10), il vient

$$a_n\left(D_1\exp\left(i\frac{h\sqrt{p_n}}{\mu}\right)+D_2\exp\left(-i\frac{h\sqrt{p_n}}{\mu}\right)\right)+$$
$$+b_n(D_1+D_2)+c_n\left(D_1\exp\left(-i\frac{h\sqrt{p_n}}{\mu}\right)+\right.$$
$$\left.+D_2\exp\left(i\frac{h\sqrt{p_n}}{\mu}\right)\right)+(a_n+b_n+c_n)\frac{f_n}{p_n}-d_n=0.$$

Pour que cette égalité soit valable quels que soient D_1 et D_2, il faut et il suffit que les coefficients de D_1 et D_2 et le terme constant soient nuls. Annulons-les:

$$a_n\exp\left(i\frac{h\sqrt{p_n}}{\mu}\right)+b_n+c_n\exp\left(-i\frac{h\sqrt{p_n}}{\mu}\right)=0,$$
$$a_n\exp\left(-i\frac{h\sqrt{p_n}}{\mu}\right)+b_n+c_n\exp\left(i\frac{h\sqrt{p_n}}{\mu}\right)=0, \quad (11)$$
$$(a_n+b_n+c_n)\frac{f_n}{p_n}-d_n=0.$$

En posant $a_n=1$, on obtient

$$c_n=1,\quad b_n=-2\cos\frac{h\sqrt{p_n}}{\mu},\quad d_n=\left(2-2\cos\frac{h\sqrt{p_n}}{\mu}\right)\frac{f_n}{p_n}.$$

La solution générale du système (11) est proportionnelle à la solution particulière obtenue. Multiplions par μ^2h^{-2} tous les coefficients a_n, b_n, c_n, d_n. Nous trouvons le schéma

$$\mu^2\frac{y_{n+1}-2\cos\frac{h\sqrt{p_n}}{\mu}y_n+y_{n-1}}{h^2}-\left(2-2\cos\frac{h\sqrt{p_n}}{\mu}\right)\frac{\mu^2 f_n}{h^2 p_n}=0. \quad (12)$$

Prenons cette relation pour l'équation aux différences correspondant au point x_n; cette normalisation du schéma (12) paraît la plus naturelle: en substituant $y(x_n)$ à y_n dans (12) et en faisant tendre h vers zéro pour x_n fixe, nous obtenons à la limite l'équation différentielle initiale

$$\mu^2 y''-p(x_n)y-f(x_n)=0.$$

Il est d'usage de tester la qualité de nouveaux programmes standard sur des problèmes types. C'est ce qu'on a voulu faire dans le cas d'un programme usuel de résolution par la méthode de Störmer du problème de Cauchy pour un système d'équations différentielles d'ordre 2; comme test on a choisi d'établir la date des éclipses de Soleil pour plusieurs siècles à venir en un point donné de la surface terrestre. On a choisi un référentiel fixe lié à la Terre par rapport auquel on a écrit les équations différentielles du mouvement de la Lune et du Soleil. Après avoir effectué l'intégration numérique on a constaté que les résultats contredisaient les valeurs réunies dans les tables d'éclipses de Soleil confectionnées il y a plusieurs siècles, alors qu'on ignorait le calcul différentiel, sans parler des machines à calculer électroniques. Face à une telle situation il faut répéter les calculs avec un pas temporel différent. Ce faisant on a constaté que les résultats différaient sensiblement d'un pas à l'autre, ce qui dénotait une erreur de calcul importante. Certains l'imputaient à la qualité médiocre des programmes standard qui n'en étaient cependant que partiellement responsables. En fait le problème proposé n'appartenait pas à la classe pour laquelle a été conçu le programme utilisé. Dans le système de coordonnées choisi, la Lune et le Soleil effectuent environ une révolution par jour, et c'est par conséquent le jour qui doit constituer la mesure caractéristique de la variation de la solution. L'erreur résultant des méthodes usuelles de résolution des problèmes ne sera petite que si le pas est de beaucoup inférieur à cette mesure caractéristique (soit à un jour). Un siècle compte environ 36 000 jours, c'est pourquoi le nombre de pas nécessaires pour calculer les éclipses un siècle durant doit être de l'ordre de centaines de milliers. Nous verrons plus loin que l'erreur de calcul dans la méthode de Störmer peut s'accumuler comme n^2. Aussi l'erreur de calcul relative sur le résultat sera en gros 10^{102-t} avec t la longueur des nombres machine. Telle est la cause de l'erreur importante sur la solution. Ce problème se résout de façon satisfaisante en construisant des schémas aux différences d'un type spécial ou en intégrant dans un système de coordonnées déterminé.

On améliore la précision de l'intégration des solutions oscillantes et à variation brusque par les techniques suivantes.

Soit une équation intégrable sous forme explicite dont les solutions se comportent comme celles de l'équation initiale. Trouvons des fonctions de ces premières solutions, qui soient fonctions régulières d'une variable indépendante. Acceptons-les ensuite en qualité de nouvelles inconnues et intégrons numériquement le système obtenu.

Soit

$$\mu^2 x'' + p(t)\, x = 0. \tag{13}$$

Lorsque $p = \text{const}$, on a

$$\mu^2 (x')^2 + p x^2 = C_0, \quad \operatorname{arctg}(x\sqrt{p}/\mu x') = C_1 t + C_2.$$

Cela permet d'introduire les nouvelles variables

$$\rho = \sqrt{\mu^2 (x')^2 + p(t)\, x^2}, \qquad \theta = \operatorname{arctg}(x\sqrt{p(t)}/\mu x').$$

Pour μ petit, les solutions de (13) vérifient les relations asymptotiques

$$x(t) \sim \frac{C}{\sqrt[4]{p(t)}} \sin\left(\frac{1}{\mu}\int_c^t \sqrt{p(\tau)}\, d\tau\right),$$

$$x'(t) \sim -\frac{C\sqrt[4]{p(t)}}{\mu} \cos\left(\frac{1}{\mu}\int_c^t \sqrt{p(\tau)}\, d\tau\right);$$

d'où

$$\frac{\mu^2 (x')^2}{\sqrt{p(t)}} + \sqrt{p(t)}\, x^2 \sim C^2, \quad \theta = \operatorname{arctg}\left(\frac{x\sqrt{p(t)}}{\mu x'}\right) \sim -\frac{1}{\mu}\int_c^t \sqrt{p(\tau)}\, d\tau.$$

Compte tenu de ces relations on fait intervenir les nouvelles inconnues

$$\rho_1 = \sqrt{\frac{\mu^2 (x')^2}{\sqrt{p(t)}} + \sqrt{p(t)}\, x^2} \quad \text{et} \quad \theta$$

et la nouvelle variable indépendante $\varphi = \int_c^t \sqrt{p(\tau)}\, d\tau$.

Les calculs d'essai permettent de définir le meilleur changement d'inconnues.

L'idée d'une autre méthode consiste à calculer à chaque pas d'intégration la perturbation de la solution d'un système simplifié, qui passe par un point déjà trouvé.

Soit par exemple le système d'équations

$$\frac{dx}{dt} = y, \quad \frac{dy}{dt} + k\frac{x}{r^3} + \varepsilon \| y \| y = 0, \quad \| x \| = r, \ \varepsilon \text{ faible.}$$

Supposons connues les valeurs approchées x^n, y^n des quantités $x(t_n)$, $y(t_n)$ et soit $x^n(t)$, $y^n(t)$ une solution du système simplifié

$$\frac{dx}{dt} = y, \quad \frac{dy}{dt} + k\frac{x}{r^3} = 0$$

avec les conditions initiales x^n, y^n. Ce système est intégré explicitement. Introduisons les nouvelles variables

$$\eta(t) = x_n(t) - x^n(t), \quad \gamma(t) = y_n(t) - y^n(t),$$

où $x_n(t)$, $y_n(t)$ est la solution exacte du système initial avec les conditions initiales x^n, y^n. Nous trouvons un système d'équations différentielles par rapport aux variables $\eta(t)$, $\gamma(t)$. Après avoir effectué un pas d'intégration par un procédé standard avec les conditions initiales $\eta(t_n) = 0$, $\gamma(t_n) = 0$ nous obtenons les valeurs approchées η^{n+1}, γ^{n+1} des quantités $\eta(t_{n+1})$, $\gamma(t_{n+1})$. Posons $x^{n+1} = x^n(t_{n+1}) + \eta^{n+1}$, $y^{n+1} = y^n(t_{n+1}) + \gamma^{n+1}$.

Au paragraphe 4 nous avons examiné l'équation (4.13) dont la solution devenait infinie lorsque $p(x) < 0$. Néanmoins, en calculant cette solution par les formules (4.9) on a vu que, selon l'estimation (4.19), l'erreur sur la solution approchée ayant traversé de tels points était faible. Les formules de calcul (4.9) sont un cas intéressant de schémas qui sont vérifiés avec une précision élevée pour les singularités de la solution de la forme $(x - a)^{-1}$.

PROBLÈME. Comment travaille le schéma (4.9) en approchant la singularité de la solution pour $p(x) \equiv \text{const} > 0$ lorsque l'équation (4.13) est intégrable sous forme explicite? Sur quelles solutions est-il exact?

§ 10. Recherche des valeurs propres par les méthodes des différences finies

Traitons un problème de valeurs propres parmi les plus simples, à savoir

$$y'' - p(x)\, y = \lambda \rho(x)\, y, \quad y(0) = 0, \quad y(X) = 0. \tag{1}$$

Donnons-nous un pas $h = XN^{-1}$ et écrivons le problème discrétisé

$$\frac{y_{n+1} - 2y_n + y_{n-1}}{h^2} - p_n y_n = \lambda \rho_n y_n, \tag{2}$$

$$n = 1, \ldots, N-1, \quad y_0 = y_N = 0; \quad p_n = p(x_n), \quad \rho_n = \rho(x_n).$$

Il est naturel d'appeler valeurs propres du problème discrétisé les valeurs λ pour lesquelles le système (2) admet une solution $y_0, \ldots, y_N$ non triviale. Supposons les valeurs propres des problèmes (1), (2) écrites dans l'ordre de décroissance

$$\lambda_1 \geqslant \lambda_2 \geqslant \ldots,$$
$$\lambda_1^h \geqslant \lambda_2^h \geqslant \ldots$$

Considérons l'exemple modèle $p(x) \equiv 0$, $\rho(x) \equiv 1$; le problème (1) se récrit alors

$$y'' = \lambda y, \quad y(0) = y(X) = 0.$$

Ses fonctions propres sont $w^m(x) = \sin\left(\frac{\pi m x}{X}\right)$ et ses valeurs propres associées $\lambda_m = -\left(\frac{\pi m}{X}\right)^2$.

En ce qui concerne le problème discret (2) qui se représente comme

$$\frac{y_{n+1} - 2y_n + y_{n-1}}{h^2} - \lambda y_n = 0, \quad y_0 = y_N = 0,$$

raisonnons comme suit: la solution générale de l'équation aux différences

$$y_{n+1} - (2 + \lambda h^2)\, y_n + y_{n-1} = 0$$

s'écrit

$$y_n = C_1 \mu_1^n + C_2 \mu_2^n,$$

avec μ_1, μ_2 racines de l'équation caractéristique

$$\mu^2 - (2 + \lambda h^2)\, \mu + 1 = 0; \tag{3}$$

évidemment $\mu_2 = \mu_1^{-1}$, donc $y_n = C_1 \mu_1^n + C_1 \mu_1^{-n}$. Les conditions $y_0 = y_N = 0$ donnent le système d'équations $C_1 + C_2 = 0$, $C_1 \mu_1^N + C_2 \mu_1^{-N} = 0$; ce système possède une solution non nulle si

son déterminant $\mu_1^{-N} - \mu_1^N = 0$; d'où

$$\mu_1 = \exp\left(\frac{2\pi i m}{2N}\right), \quad m = \ldots, \ -1, \ 0, \ 1, \ \ldots$$

En se servant de (3) on exprime les valeurs λ^h par les valeurs μ_1

$$\lambda_m^h = \frac{\mu_1 + \mu_1^{-1} - 2}{h^2} = \frac{\exp\left(\frac{\pi i m}{N}\right) + \exp\left(-\frac{\pi i m}{N}\right) - 2}{h^2} =$$

$$= \frac{N^2}{X^2}\left(2\cos\frac{\pi m}{N} - 2\right) = -4\frac{N^2}{X^2}\sin^2\frac{\pi m}{2N}.$$

Prenons, pour fixer les idées, $C_1 = (2i)^{-1}$; les fonctions propres correspondantes ont alors la forme

$$W_n^m = \frac{1}{2i}\left(\exp\left(\frac{\pi i m n}{N}\right) - \exp\left(-\frac{\pi i m n}{N}\right)\right) = \sin\frac{\pi m n}{N}.$$

Les valeurs propres $\lambda_1^h, \ \ldots, \ \lambda_{N-1}^h$ étant distinctes, il en est de même des fonctions propres associées $\sin\frac{\pi n}{N}, \ \ldots, \ \sin\frac{\pi (N-1) n}{N}$. Comme le problème (2) est un problème de valeurs propres pour une matrice de dimension $N - 1$, nous obtenons un système complet de fonctions propres; chacune des W_n^m est, lorsque $m \leqslant 0$ ou $m \geqslant N$, identiquement nulle ou proportionnelle à l'une des fonctions $W_n^1, \ \ldots \ \ldots, \ W_n^{N-1}$ énumérées. Nous avons choisi par hasard un exemple où les nœuds $x_n = nh$ du réseau vérifient

$$W_n^m = w^m(nh).$$

Cette égalité n'est pas vraie dans le cas général; cependant, la nature de la proximité des valeurs propres de ces problèmes est également caractéristique du cas général. Puisque $\cos x = 1 - \frac{x^2}{2} + (\cos\theta x)\frac{x^4}{24}$ avec $|\theta| \leqslant 1$, on a

$$\lambda_m^h = -\frac{N^2}{X^2}\left(\frac{\pi^2 m^2}{N^2} - \frac{\cos\theta_m}{12}\left(\frac{\pi m}{N}\right)^4\right),$$

ou bien

$$\lambda_m^h = -\left(\frac{\pi m}{X}\right)^2 + \frac{\cos\theta_m}{12}\left(\frac{\pi m}{X}\right)^4 h^2. \tag{4}$$

Cette formule montre que $\lambda_m^h - \lambda_m = O(h^2)$ pour m fixe; d'autre part, l'erreur absolue et l'erreur relative croissent de façon monotone avec m et, par exemple,

$$\frac{\lambda_{N-1}}{\lambda_{N-1}^h} = \frac{-\pi^2 (N-1)^2}{-4N^2 \sin^2\frac{\pi (N-1)}{2N}} \sim \frac{\pi^2}{4}.$$

L'égalité (4) peut s'écrire sous forme de l'estimation $|\lambda_m^h - \lambda_m| \leqslant \leqslant C\lambda_m^2 h^2$, avec C indépendant de λ_m et h. Dans le cas de $p(x)$, $\rho(x)$ deux fois continûment dérivables, on peut obtenir la même estimation.

Indiquons comment on peut obtenir une estimation du type

$$|\lambda_m^h - \lambda_m| \leqslant C_0(\lambda_m) h^2. \tag{5}$$

Désignons par $w^\lambda(x)$ la solution du problème de Cauchy

$$w'' - p(x) w = \lambda\rho(x) w, \quad w(0) = 0, \ w'(0) = 1$$

et par w_n^λ la solution du problème de Cauchy discret $w_0^\lambda = 0$, $w_1^\lambda = h$,

$$\frac{w_{n+1}^\lambda - 2w_n^\lambda + w_{n-1}^\lambda}{h^2} - p_n w_n^\lambda = \lambda\rho_n w_n^\lambda.$$

Les zéros de la fonction $g(\lambda) = w^\lambda(X)$ sont évidemment les valeurs propres du problème différentiel et les zéros de $g_h(\lambda) = w_N^\lambda$ celles du problème sous forme discrète. Les évaluations de l'erreur sur la solution de ce problème de Cauchy (§ 2) font que

$$|g_h(\lambda) - g(\lambda)| \leqslant C_1(\lambda) h^2.$$

On démontre à l'aide d'une technique standard que $g'(\lambda) \neq 0$ si $g(\lambda) = 0$; aussi la fonction $g(\lambda)$ change-t-elle de signe dans le voisinage de chacun de ses zéros.

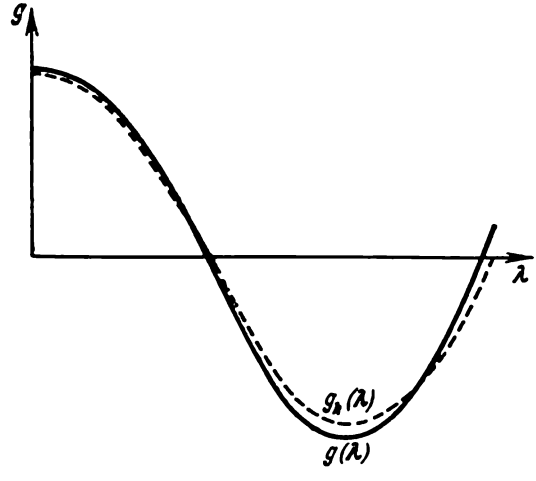


Fig. 9.10.1

La fonction $g_h(\lambda)$ (fig. 9.10.1) change elle aussi de signe à proximité des zéros de $g(\lambda)$ et ne s'annule pas loin de ses racines; c'est pourquoi les zéros de $g_h(\lambda)$ et de $g(\lambda)$ sont voisins, et le degré de leur proximité est évalué par (5) de la manière suivante. Soient λ^h et λ deux racines voisines de $g_h(\lambda)$

et $g(\lambda)$. Selon la dernière inégalité $|g(\lambda^h)| \leqslant C_1(\lambda^h)h^2$; cette inégalité s'écrit

$$|g(\lambda^h) - g(\lambda)| \leqslant C_1(\lambda^h)h^2$$

ou encore

$$|g'(\bar{\lambda})(\lambda^h - \lambda)| \leqslant C_1(\lambda^h)h^2;$$

il en résulte

$$|\lambda^h - \lambda| \leqslant \frac{C_1(\lambda^h)}{|g'(\bar{\lambda})|}h^2.$$

On minore ensuite $|g'(\lambda)|$ dans le domaine avec $|g(\lambda)| \leqslant C_1(\lambda^h)h^2$, puis on démontre que le voisinage de chaque zéro de $g(\lambda)$ ne contient qu'une seule racine de $g_h(\lambda)$.

Afin de résoudre (1) avec une précision meilleure, on utilise des approximations aux différences plus exactes quelconques de l'équation $y'' - q(x)y = 0$. Traitons un exemple.

Nous avons construit au paragraphe 1 un schéma aux différences qui approchait la dernière équation avec l'erreur $O(h^4)$:

$$\frac{\delta^2 y_n}{h^2} - q_n y_n - \frac{1}{12}\delta^2(q_n y_n) = 0 \tag{6}$$

(nous avons légèrement modifié les notations). L'équation (1) se met sous la forme considérée lorsque $q(x) = p(x) + \lambda\rho(x)$. D'où le problème discret de valeurs propres

$$\frac{\delta^2 y_n}{h^2} - (p_n + \lambda\rho_n)y_n - \frac{1}{12}\delta^2((p_n + \lambda\rho_n)y_n) = 0, \tag{7}$$

$$n = 1; \ldots, N-1, \quad y_0 = y_N = 0.$$

On montre que les valeurs propres de ce problème satisfont à l'estimation

$$\lambda_m^h - \lambda_m = O(\lambda_m^3 h^4).$$

Le système d'équations (7) s'écrit $A\mathbf{y} - \lambda B\mathbf{y} = 0$, écriture formellement plus compliquée que (2) car la matrice B n'est pas diagonale. En élevant la précision on voit apparaître des problèmes discrets qui ont de prime abord un aspect encore plus complexe, à l'image de l'exemple suivant : on désire trouver λ tel que le système

$$\alpha_n(\lambda, h)y_{n-1} - \beta_n(\lambda, h)y_n + \gamma_n(\lambda, h)y_{n+1} = 0, \tag{8}$$

$$n = 1, \ldots, N-1, \quad y_0 = y_N = 0,$$

admet une solution y_n non nulle.

Prenons le cas $\gamma_n \neq 0$, fixons un certain λ, $w_1 \neq 0$ (par exemple $w_1 = h$) et déterminons de proche en proche $w_2^\lambda, \ldots, w_N^\lambda$ moyennant la relation

$$\alpha_n(\lambda, h)w_{n-1}^\lambda - \beta_n(\lambda, h)w_n^\lambda + \gamma_n(\lambda, h)w_{n+1}^\lambda = 0. \tag{9}$$

Si $w_N^\lambda = 0$, λ est une valeur propre et w_n^λ une fonction propre; si $w_N^\lambda \neq 0$, ce λ n'est pas valeur propre du problème (8).

On recherche les valeurs propres de (8) se confondant avec les zéros de w_N^λ par une méthode itérative qui fournit les zéros d'une fonction à partir de ses valeurs. Ce processus est parfois facilité: en effet si la fonction w_n^λ présente sur $(0, X)$ j changements de signe, alors $\lambda_j < \lambda < \lambda_{j+1}$.

En principe les formules de récurrence (9) sont les plus qualifiées pour le calcul des valeurs de w_N^λ pour les λ divers.

L'algorithme de calcul des valeurs de w_N^λ que nous proposons coïncide avec le procédé de résolution du problème de Cauchy discret pour (1). Si les solutions du problème de Cauchy différentiel augmentent fortement, on constate comme plus haut des erreurs de calcul importantes.

Les valeurs propres s'obtiennent également par balayage. Passons sur le principe de la méthode et bornons-nous à écrire les formules de calcul. Posons $\frac{w_n^\lambda}{w_{n+1}^\lambda} = C_n$; la relation (9) s'écrit alors $\alpha_n C_{n-1} C_n - \beta_n C_n + \gamma_n = 0$, d'où

$$C_n = \frac{\gamma_n}{\beta_n - \alpha_n C_{n-1}}. \tag{10}$$

Si w_n^λ et w_{n+1}^λ sont du même signe, C_n est supérieur à zéro, dans le cas contraire on a $C_n < 0$. C'est pourquoi les changements de signe de C_n nous permettent de déterminer le nombre de changements de signe de w_n^λ. Le coefficient C_n peut être, nous l'avons vu au § 4, très important si bien que la méthode en question s'applique le plus souvent à la recherche de la première valeur propre.

Voyons encore un procédé pour ramener un problème aux limites de valeurs propres à la recherche des zéros d'une certaine fonction. Soit le problème

$$\mathbf{y}' - A(x)\,\mathbf{y} - \lambda C(x)\,\mathbf{y} = 0, \quad B\mathbf{y}(0) = 0, \quad D\mathbf{y}(X) = 0. \tag{11}$$

En se donnant λ et en réalisant le balayage orthogonal on détermine successivement les matrices $G_s(X_s)$, $s = 1, \ldots, m$. Si le système d'équations (cf. § 6)

$$D(G_m(X_m)\,\beta^m) = 0$$

admet une solution non nulle, il en est de même du problème de départ (11) et la valeur λ correspondante est bien une valeur propre. Ainsi, la recherche d'une valeur propre équivaut en l'occurrence à trouver une racine de l'équation

$$\det(DG_m(X_m)) = 0.$$

§ 11. Optimisation de la répartition des points d'intégration

La régularité des solutions des équations et systèmes différentiels peut varier d'un tronçon du segment d'intégration à l'autre. En évaluant l'erreur dans la méthode de Runge-Kutta nous avons vu que la contribution de l'erreur d'intégration sur un pas $[x_j, x_{j+1}]$ à l'erreur totale était proportionnelle au point $x_N = x_0 + X$ au facteur $\exp\left(\int_{x_j}^{x_N} f_y(x, y(x))\,dx\right)$ qui dépend de j. On voit donc se poser, pour certaines classes d'équations différentielles, la question de l'optimisation de la répartition des points d'intégration. Supposons, pour simplifier, que la condition initiale est donnée exactement et qu'il n'y a pas d'arrondi. L'erreur sur le résultat de l'intégration numérique par la technique de Runge-Kutta au point x_N ne dépasse pas, selon (8.4.6),

$$S_N = \sum_{j=1}^{N} |y_j - y_{j-1}(x_j)| \exp\left(\int_{x_j}^{x_N} f_y(x, \tilde{y}_j(x))\,dx\right). \qquad (1)$$

Soit $x_j = \varphi\left(\frac{j}{N}\right)$, $\varphi(0) = x_0$, $\varphi(1) = x_0 + X$, $\varphi(t)$ une fonction régulière. Conformément à la formule de Lagrange

$$x_j - x_{j-1} = \varphi\left(\frac{j}{N}\right) - \varphi\left(\frac{j-1}{N}\right) = \frac{1}{N}\varphi'(\bar{t}_j),$$

où $\frac{j-1}{N} \leqslant \bar{t}_j \leqslant \frac{j}{N}$, et donc

$$H = \max_{0<j\leqslant N}(x_j - x_{j-1}) \leqslant \frac{1}{N}\max_{[0,\,1]}|\varphi'(t)|, \quad H \to 0 \text{ pour } N \to \infty;$$

l'estimation du § 4, chap. VIII entraîne, par conséquent,

$$\max_{x_0 \leqslant x_j \leqslant x \leqslant x_0 + X} |y_j(x) - y(x)| \to 0$$

pour N tendant vers l'infini. On écrit par suite les égalités asymptotiques

$$\exp\left(\int_{x_j}^{x_N} f_y(x, \tilde{y}_j(x))\,dx\right) \sim \exp\left(\int_{x_j}^{x_N} f_y(x, y(x))\,dx\right),$$

$$y_j - y_{j-1}(x_j) \sim \psi(x_j, y(x_j))(x_j - x_{j-1})^{k+1} \sim$$

$$\sim \psi\left(\varphi\left(\frac{j}{N}\right),\ y\left(\varphi\left(\frac{j}{N}\right)\right)\right)\left(\frac{1}{N}\varphi'\left(\frac{j}{N}\right)\right)^{k+1}.$$

L'expression (1) de S_N se met alors sous la forme

$$S_N \sim \frac{1}{N^k} \sum_{j=1}^{N} \frac{1}{N} \Phi\left(\frac{j}{N}\right), \tag{2}$$

où

$$\Phi(t) = |\psi(\varphi(t), y(\varphi(t)))| (\varphi'(t))^{k+1} \exp\Big(\int_{\varphi(t)}^{x_0+X} f_y(x, y(x))\, dx \Big). \tag{3}$$

Lorsque $\Phi(t)$ est régulière et N tend vers l'infini, la quantité $\sum_{j=1}^{N} \frac{1}{N} \Phi\left(\frac{j}{N}\right)$ a pour limite

$$I = \int_0^1 \Phi(t)\, dt \tag{4}$$

et donc

$$S_N \sim N^{-k} \int_0^1 \Phi(t)\, dt.$$

Prenons φ pour nouvelle variable dans l'intégrale I. Celle-ci s'écrit alors

$$I = \int_{x_0}^{x_0+X} L(\varphi)\, d\varphi,$$

avec

$$L(\varphi) = |\psi(\varphi, y(\varphi))| (t'(\varphi))^{-k} \exp\Big(\int_{\varphi}^{x_0+X} f_y(x, y(x))\, dx \Big). \tag{5}$$

Minimiser cette intégrale sous la forme (4) grâce à un choix convenable de $\varphi(t)$ équivaut à la minimiser sous la forme (5) en choisissant cette fois la fonction inverse $t(\varphi)$. Vu l'égalité $\frac{\partial L}{\partial t} = 0$, l'équation d'Euler

$$\frac{d}{d\varphi}\left(\frac{\partial L}{\partial t'}\right) - \frac{\partial L}{\partial t} = 0$$

s'écrit dans ce cas $\frac{d}{d\varphi}\left(\frac{\partial L}{\partial t'}\right) = 0$; il en résulte

$$\frac{\partial L}{\partial t'} = -k |\psi(\varphi, y(\varphi))| (t'(\varphi))^{-k-1} \exp\Big(\int_{\varphi}^{x_0+X} f_y(x, y(x))\, dx \Big) = \text{const.}$$

On obtient en revenant à la variable φ :

$$(\varphi'(t))^{k+1} \,|\, \psi(\varphi, y(\varphi))\,| \exp\Big(\int_{\varphi}^{x_0+X} f_y(x, y(x))\,dx\Big) = C_1. \qquad (6)$$

La solution de cette équation différentielle dépend de C_1 et d'une constante C_2 qui se déduisent à partir des conditions aux limites $\varphi(0) = x_0$, $\varphi(1) = x_0 + X$. On légitime ces constructions pour $\psi(\varphi, y(\varphi))$ ou

$$\exp\Big(\int_{\varphi}^{x_0+X} f_y(x, y(x))\,dx\Big)$$

non bornés ou non réguliers par les mêmes considérations que dans le problème d'intégration numérique. D'autre part, ce que nous avons dit sur l'utilisation pratique de la relation (6) reste en vigueur pour ce dernier problème.

Indiquons une circonstance non évidente. L'équation (6) peut s'écrire

$$(\varphi'(t))^{k+1} \,|\, \psi(\varphi, y(\varphi))\,| \exp\Big(-\int_{0}^{\varphi} f_y(x, y(x))\,dx\Big) = C, \qquad (7)$$

où ne figurent ni le point initial, ni le point final d'intégration. Cette équation est vérifiée par $\varphi(t)$ et $\varphi(at + b)$. En nous donnant un certain N, intégrons simultanément l'équation de départ et (7) en choisissant à chaque fois les pas à partir de la condition $x_j - x_{j-1} \sim \frac{1}{N}\varphi'(t_{j-1})$. Il est très probable que nous arrivons au point final $x_0 + X$ avec un nombre de points N_1 différent de N et cette répartition ne sera pas alors la répartition optimale correspondant à N donné. Ce qui est naturel puisqu'en démarrant les calculs nous ne pouvons rien dire sur le comportement futur de la solution ni sur la valeur de C à prendre dans le second membre de (7). On montre néanmoins que par suite des propriétés indiquées de la solution de (7) $\varphi(t)$ est bien la fonction cherchée et $\varphi(j/N_1)$ la répartition optimale (à l'erreur d'intégration numérique près) correspondant au nombre N_1 de points.

La nécessité de calculer les valeurs de la fonction $\psi(x, y(x))$ entrave l'intégration immédiate des équations (1) et (7). On ferait mieux, au lieu de chercher directement cette fonction, d'utiliser le contrôle de l'erreur par pas. En intégrant numériquement (7) on tiendra également compte du caractère asymptotique de ces relations. Au voisinage des points en lesquels $\psi(x, y(x)) = 0$, ce sont les termes en $(x_j - x_{j-1})^{k+2}$ qui deviennent essentiels dans le reste.

L'intégration numérique complémentaire de l'équation (7) risque de compliquer singulièrement la résolution du problème. Aussi pour optimiser la répartition des points opère-t-on souvent de la sorte. Soit à résoudre des problèmes d'une classe donnée. Traitons un problème qui soit un modèle pour cette classe et tel que l'équation (7) admette une solution explicite. Etablissons, sur l'exemple de ce problème, entre le pas (ou la mesure de l'erreur par pas (cf. § 3, chap. VIII)) et l'allure de la solution une relation telle que la répartition des points soit voisine de l'optimum. Intégrons ensuite tous les problèmes de la classe avec un pas correspondant à cette relation.

Dans d'autres cas, une certaine forme de cette relation est donnée d'avance. Soit un système d'équations d'ordre m que l'on intègre avec vérification de l'erreur par pas d'après les formules (8.3.2). S'il s'agit d'un système d'équations, le terme de contrôle r est un vecteur $\mathbf{r}_n = (r_n^1, \ldots, r_n^m)$. Dans plusieurs programmes le pas d'intégration est choisi à partir de la condition

$$\| \mathbf{r}_n \| / (\max (M, \| \mathbf{y}_n \|)) \approx \varepsilon = \text{const},$$

ou à partir de la condition suivante qui est plus souple,

$$\max_{1 \leqslant k \leqslant m} \frac{| r_n^k |}{\max (| M_k |, | (y_n)_k |)} \approx \varepsilon = \text{const}.$$

Les paramètres M, M_k sont choisis à partir de l'optimalité de la répartition des points dans le problème proposé.

Voici quelques exemples d'optimisation.

Supposons que le problème de Cauchy $y' = My$ avec la condition initiale $y(0)$ se résout par la méthode d'Euler. Alors

$$\psi (x, y(x)) = -\frac{1}{2} y''(x) = -\frac{M^2}{2} y(0) \exp (Mx)$$

et l'équation (6) s'écrit

$$(\varphi_i')^2 \frac{M^2}{2} | y(0) | \exp (M\varphi) \exp (M(X - \varphi)) = \text{const}.$$

D'où $\varphi_i' = \text{const}$, i. e. il faut prendre une répartition uniforme.

PROBLÈME. Supposons que la mesure de l'erreur par pas ρ s'obtient (de même qu'au paragraphe 3 du chap. VIII) comme la différence entre une valeur approchée au bout d'un pas double et une valeur après deux pas identiques. Vérifier que la répartition est uniforme, i.e. optimale, si

$$\left| \frac{r_n}{y_n} \right| \approx \varepsilon = \text{const}.$$

Traitons le problème aux limites $\mu^2 y'' - y = f(x)$ sur le segment $[0, 1]$ pour $y(0)$, $y(1)$ donnés et μ faible. Il y a intérêt à étudier les méthodes de résolution de ce problème en tant que modèle de problèmes avec *couche limite*. Sa solution pour $f(x) \equiv 0$ est

$$y(x) = \frac{y(0) \left(\exp \left(-\frac{x}{\mu} \right) - \exp \left(\frac{x-2}{\mu} \right) \right)}{1 - \exp \left(-\frac{2}{\mu} \right)} + \\ + \frac{y(1) \left(\exp \left(\frac{x-1}{\mu} \right) - \exp \left(-\frac{x+1}{\mu} \right) \right)}{1 - \exp \left(-\frac{2}{\mu} \right)}.$$

En dérivant directement on voit que

$$| y^{(k)}(x) | = O \left(\frac{1}{\mu^k} y(0) \exp \left(-\frac{x}{\mu} \right) \right) + O \left(\frac{1}{\mu^k} y(1) \exp \left(\frac{x-1}{\mu} \right) \right).$$

Lorsque $f(x) \not\equiv 0$, cette estimation devient

$$| y^{(k)}(x) | = O \left(\frac{1}{\mu^k} \exp \left(-\frac{x}{\mu} \right) \right) + O \left(\frac{1}{\mu^k} \exp \left(\frac{x-1}{\mu} \right) \right) + O(1). \qquad (8)$$

Ainsi, les propriétés de dérivabilité de la solution empirent brusquement pour μ petits au voisinage des points $x = 0$ et $x = 1$, et on appliquera donc dans ce voisinage un pas d'intégration plus petit. Donnons-nous les points d'intégration $x_j = \varphi\left(\frac{j}{N}\right)$, $\varphi(0) = 0$, $\varphi(1) = 1$, et remplaçons la dérivée $y''(x_j)$ par la différence divisée double $y''(x_j) \approx 2y(x_{j-1}; x_j; x_{j+1})$. Nous aboutissons au système d'équations aux différences

$$l(y_j) = 2\mu^2 \frac{\dfrac{y_{j+1}-y_j}{x_{j+1}-x_j} - \dfrac{y_j-y_{j-1}}{x_j-x_{j-1}}}{x_{j+1}-x_{j-1}} - y_j = f_j, \qquad (9)$$

$$j = 1, \ldots, N-1, \quad y_0 = y(0), \quad y_N = y(1).$$

En exprimant la différence divisée moyennant la dérivée il vient

$$2y(x_{j-1}; x_j; x_{j+1}) = y''(\zeta_j),$$

avec $x_{j-1} \leqslant \zeta_j \leqslant x_{j+1}$. Aussi

$$r_n = l(y(x_n)) - f_n = \mu^2 (y''(\zeta_n) - y''(x_n)) = \mu^2 (\zeta_n - x_n)\, y^{(3)}(\eta_n). \qquad (10)$$

D'où l'inégalité

$$|r_n| \leqslant \sigma_n = \mu^2 (\max(|x_{n+1}-x_n|, |x_n-x_{n-1}|))\left(\max_{[x_{n-1}, x_{n+1}]} |y^{(3)}(x)|\right). \qquad (11)$$

Nous trouvons à l'aide de (9) et (10)

$$l(R_j) = -r_j, \quad j = 1, \ldots, N-1, \quad R_0 = R_N = 0, \qquad (12)$$

donnant l'erreur sur la solution approchée $R_n = y_n - y(x_n)$. Il faut ensuite obtenir une représentation asymptotique de l'erreur, puis poser et résoudre le problème asymptotique correspondant du calcul des variations.

Si nous voulons minimiser le maximum de l'erreur non pas en des points isolés mais sur le segment tout entier, nous aurons à effectuer de longs calculs. C'est pourquoi nous remplaçons le problème d'optimisation par un autre, plus simple. Le lecteur est prié de vérifier que

$$\max_n |R_n| \leqslant \max_n |r_n|.$$

Il s'ensuit

$$\max_n |R_n| \leqslant \max_n |\sigma_n|.$$

Soit à minimiser le second membre de cette inégalité. Lorsque $N \to \infty$, on a

$$\sigma_n \sim \frac{\mu^2}{N} \varphi'(n/N) \left| y''\left(\varphi\left(\frac{n}{N}\right)\right)\right|.$$

Donc, ce problème asymptotique revient à minimiser

$$z(\varphi) = \max_{[0, 1]} \varphi'(t)\, |y^{(3)}(\varphi)|$$

dans les conditions $\varphi(0) = 0$ et $\varphi(1) = 1$.

Montrons que la borne inférieure de $z(\varphi)$, égale à

$$z_0 = \left(\int_0^1 \frac{dt}{|y^{(3)}(\varphi(t))|}\right)^{-1},$$

est atteinte pour $\varphi'(t) = (|y^{(3)}(\varphi(t))|\, z_0)^{-1}$. Supposons, en effet, qu'une fonction $\varphi(t)$ satisfait, avec t quelconque, à l'inégalité $\varphi'(t)\, |y^{(3)}(\varphi(t))| < z_0$;

il en découle

$$\varphi'(t) < |\, y^{(3)}(\varphi(t))\,|^{-1} z_0.$$

En intégrant de 0 à 1, il vient

$$1 = \varphi(1) - \varphi(0) < \int_0^1 |\, y^{(3)}(\varphi(t))\,|^{-1} z_0\, dt = 1,$$

d'où contradiction.

Pour ce $\varphi'(t)$ on a

$$\mu\varphi'(t)\,|\, y^{(3)}(\varphi(t))\,| = \mu z_0^{-1} = \text{const}. \tag{13}$$

Dans le cas considéré, selon (8),

$$|\, y^{(3)}(t)\,| \leqslant c\left(\frac{1}{\mu^3}\left(\exp\left(-\frac{x}{\mu}\right)+\exp\left(\frac{x-1}{\mu}\right)\right)+1\right).$$

Considérons donc, au lieu de (13), l'équation

$$\varphi'(t)\left(\frac{1}{\mu^3}\left(\exp\left(-\frac{\varphi}{\mu}\right)+\exp\left(\frac{\varphi-1}{\mu}\right)\right)+1\right) = \text{const}.$$

Cette équation et les conditions aux limites $\varphi(0) = 0$, $\varphi(1) = 1$ déterminent de façon unique la fonction $\varphi(t)$.

Notons qu'en réalité l'ordre d'approximation du schéma considéré est deux, ce dont il faudra tenir compte lorsqu'on choisira la répartition des points.

§ 12. Influence de l'erreur de calcul en fonction de l'écriture de l'équation aux différences finies

On sait que la croissance de l'erreur de calcul varie selon la technique de résolution des équations différentielles. On se propose maintenant d'étudier un problème particulier important : comment l'erreur de calcul dépend-elle de l'écriture de l'équation aux différences finies? Bien que les raisonnements soient faits sur l'exemple du problème de Cauchy, ils valent également pour les problèmes aux limites. Adressons-nous par exemple à la méthode d'Euler

$$y_{n+1} = y_n + hf(x_n, y_n). \tag{1}$$

Les calculs effectifs donnent les quantités y_n^* reliées par

$$y_{n+1}^* = y_n^* + hf(x_n, y_n^*) + \delta_n; \tag{2}$$

le terme δ_n est, nous l'avons dit, le résultat des erreurs commises en calculant les valeurs de la fonction $f(x_n, y_n^*)$; de l'erreur d'arrondi sur le produit $hf(x_n, y_n^*)$; de l'erreur résultant de l'addition des nombres y_n^* et de la valeur arrondie $hf(x_n, y_n^*)$.

Introduisons la notation $y_n^* - y_n = \Delta_n$; conformément à la formule de Lagrange,

$$f(x_n, y_n^*) - f(x_n, y_n) = l_n \Delta_n$$

avec $l_n = f_y(x_n, \bar{y}_n)$. Supposons qu'on a toujours $|f_y(x, y)| \leqslant L$. En retranchant (1) de (2) nous trouvons

$$\Delta_{n+1} = (1 + l_n h)\,\Delta_n + \delta_n,$$

d'où

$$|\Delta_{n+1}| \leqslant (1 + Lh)\,|\Delta_n| + \delta, \quad \delta = \max_n \delta_n. \tag{3}$$

Considérons l'équation aux différences

$$z_{n+1} = (1 + Lh)\,z_n + \delta,$$

appelée *équation majorante* de (3). Sa solution s'écrit, dans la condition initiale $z_0 = |\Delta_0|$:

$$z_n^0 = |\Delta_0|\,(1 + Lh)^n + \delta\,\frac{(1+Lh)^n - 1}{Lh}.$$

LEMME.

$$|\Delta_n| \leqslant z_n^0 \tag{4}$$

quels que soient $n \geqslant 0$.

DÉMONSTRATION. La proposition (4) est évidente pour $n = 0$. Supposons qu'elle est vraie pour un certain n. Alors

$$|\Delta_{n+1}| \leqslant (1 + Lh)\,z_n^0 + \delta = z_{n+1}^0,$$

ce qui démontre le lemme.

Si le segment d'intégration est $[x_0, x_0 + X]$, alors

$$nh \leqslant X, \quad (1 + Lh)^n \leqslant (\exp(Lh))^n \leqslant \exp(LX).$$

Selon la formule de Lagrange

$$x_2^N - x_1^N = N\bar{x}^{N-1}(x_2 - x_1),$$

et donc

$$(1 + Lh)^n - 1 = n\,(1 + \theta Lh)^{n-1} Lh, \text{ où } 0 \leqslant \theta \leqslant 1.$$

Il en résulte

$$(1 + Lh)^n - 1 \leqslant nLh\,(\exp(Lh))^{n-1} \leqslant LX \exp(LX).$$

L'estimation définitive a la forme

$$|\Delta_n| \leqslant |z_n^0| \leqslant |\Delta_0| \exp(LX) + \delta X h^{-1} \exp(LX).$$

Soit $\Delta_0 = 0$. L'erreur $\Delta_n = y_n^* - y_n$ est alors majorée, pour X fixe, par $O(\delta h^{-1})$.

On ne saurait améliorer l'ordre de cette estimation. Ainsi, lorsque $\Delta_0 = 0$, $f_y \equiv 0$, $\delta_n \equiv \delta$, on a $\Delta_{n+1} = \Delta_n + \delta$ et donc

$$\Delta_N = N\delta = \delta X h^{-1}.$$

Nous pourrions très bien, au lieu d'une évaluation rigoureuse, nous borner à des raisonnements vraisemblables, d'où il découlera que

l'erreur d'arrondi peut être une quantité en δh^{-1}. La relation

$$y^*_{n+1} = y^*_n + hf(x_n, y^*_n) + \delta_n$$

se récrit

$$y^*_{n+1} = y^*_n + hf^*(x_n, y^*_n),$$

où

$$f^*(x_n, y^*_n) = f(x_n, y^*_n) + \delta_n h^{-1}.$$

Ainsi, l'intégration numérique avec arrondi de l'équation $y' = f(x, y)$ donne le même résultat que l'intégration sans arrondi de la même équation dont le second membre prend les valeurs aux points $f^*(x_n, y^*_n)$ du réseau. Traitons le cas $\delta_n \equiv \delta$. La différence entre les solutions des équations différentielles

$$y' = f(x, y) \quad \text{et} \quad y' = f(x, y) + \delta h^{-1}$$

est de l'ordre de la différence de leurs seconds membres, i.e. δh^{-1}. Rien ne fait penser que les solutions des équations aux différences

$$y_{n+1} = y_n + hf(x_n, y_n) \quad \text{et} \quad y^*_{n+1} = y^*_n + h(f(x_n, y^*_n) + \delta h^{-1})$$

diffèrent d'une quantité d'un ordre de grandeur moindre. Si δ est en 2^{-t}, où t est la longueur des nombres en machine, la solution de l'équation aux différences varie d'une quantité de l'ordre de $2^{-t}h^{-1}$.

Nous avons admis $\delta_n \equiv \delta$. Considérons le *problème de calcul de* $\int_0^1 f(x)\,dx$, cas particulier du problème en question pour $y(0) = 0$. Soit $f(x) = 2/3$, $h = 2^{-k}$, $t-k$ impair. Lorsque $x_n > 3/4$, y_n se trouve dans l'intervalle (1/2, 1). Dans l'addition

$$\begin{array}{rl} & y_n = 0,\ 1\alpha_2 \ldots\ldots \alpha_t \\ + & hf_n = 0,\ \underbrace{0 \ldots 0}_{k}\underbrace{10 \ldots 1}_{t-k}0101 \ldots \end{array}$$

avec arrondi, chaque fois on laisse tomber alors les quantités

$$2^{-t} \cdot 0{,}0101 \ldots = (1/3)\,2^{-t}.$$

Si bien que sur ce segment $\delta_n \equiv \delta$ est effectivement de l'ordre de 2^{-t}.

Passons à l'intégration de l'équation $y'' = f(x, y)$ par la méthode simple de Störmer

$$\frac{y_{n+1} - 2y_n + y_{n-1}}{h^2} = f(x_n, y_n)$$

selon la formule de calcul

$$y_{n+1} = 2y_n - y_{n-1} + h^2 f(x_n, y_n);$$

les valeurs y^*_n obtenues réellement sont reliées par

$$y^*_{n+1} = 2y^*_n - y^*_{n-1} + h^2 f(x_n, y^*_n) + \delta_n. \qquad (5)$$

On peut considérer ces valeurs comme le résultat du calcul sans arrondi par la formule

$$y^*_{n+1} = 2y^*_n - y^*_{n-1} + h^2 f^*(x_n, y^*_n),$$

où

$$f^*(x_n, y^*_n) = f(x_n, y^*_n) + \delta_n h^{-2}.$$

Traitons le cas $\delta_n \equiv \delta$. Les solutions des équations différentielles

$$y'' = f(x, y) \quad \text{et} \quad y'' = f(x, y) + \delta h^{-2}$$

diffèrent d'une quantité en δh^{-2}. On montre que dans le cas de l'équation $y'' = \alpha$, $\alpha = \text{const}$, on a éventuellement $\delta_n \equiv \delta$ avec δ de l'ordre de 2^{-t}. Comme pour l'équation du premier ordre nous en concluons que l'erreur de calcul totale dans l'algorithme examiné peut être en $2^{-t}h^{-2}$.

Prenons le cas où une telle erreur est inadmissible. Nous pourrions écrire l'équation considérée sous forme de système d'équations du premier ordre

$$y' = v, \quad v' = f(x, y) \tag{6}$$

et faire appel à une technique d'intégration numérique. Comme signalé plus haut, il en résulterait une efficience moindre car les méthodes générales ne tiennent pas compte du caractère spécifique de ce système.

Tâchons d'écrire la formule de calcul considérée comme une formule d'intégration du système (6) et introduisons la nouvelle variable discrète

$$\frac{y_n - y_{n-1}}{h} = z_n ;$$

l'équation (5) se récrit alors

$$\frac{z_{n+1} - z_n}{h} = f(x_n, y_n) ;$$

trouvons les valeurs consécutives de y_n, z_n à l'aide du couple de formules

$$z_{n+1} = z_n + hf(x_n, y_n), \quad y_{n+1} = y_n + hz_{n+1}. \tag{7}$$

S'il y a arrondi, nous avons respectivement

$$z^*_{n+1} = z^*_n + hf(x_n, y^*_n) + \alpha_n, \quad y^*_{n+1} = y^*_n + hz^*_{n+1} + \beta_n$$

avec α_n, $\beta_n = O(2^{-t})$. Ces relations s'écrivent également

$$\begin{aligned} &z^*_{n+1} = z^*_n + hf^*(x_n, y^*_n), \quad f^*(x_n, y^*_n) = f(x_n, y^*_n) + \alpha_n h^{-1}, \\ &y^*_{n+1} = y^*_n + h(z^*_{n+1} + g(x_{n+1})), \quad g(x_{n+1}) = \beta_n h^{-1}. \end{aligned} \tag{8}$$

Si les formules (7) s'interprètent comme les formules d'intégration numérique du système $z' = f(x, y)$, $y' = z$, les relations (8) corres-

pondent au système

$$z' = f^* (x, y), \quad y' = z + g (x).$$

La différence des seconds membres de ces systèmes est une quantité de l'ordre de $O(2^{-t}h^{-1})$. C'est pourquoi il y a des chances pour que les solutions du problème aux différences avec arrondi et de celui sans arrondi diffèrent d'un nombre du même ordre de grandeur. Cette affirmation appelle, cela va sans dire, une attitude critique, car pour certains schémas aux différences finies des erreurs faibles peuvent, nous l'avons vu, perturber catastrophiquement le résultat.

Nos raisonnements sur l'influence de l'erreur de calcul dans des méthodes concrètes d'intégration des équations du premier et du deuxième ordre s'appuient uniquement sur les propriétés du schéma aux différences finies qui sont liées à l'ordre de l'équation différentielle. Ils s'étendent donc probablement à d'autres méthodes des différences finies. Ainsi, si l'on intègre l'équation $y^{(k)} = f(x, y)$ et si on utilise directement le schéma

$$\nabla^k y_n - h^k \sum_{i=0}^{l} a_{-i} f(x_{n-i}, y_{n-i}) = 0,$$

l'erreur de calcul sera probablement de l'ordre de $2^{-t}h^{-k}$. Aussi, dans le cas d'équations d'ordre élevé, il est encore plus important de transformer le schéma de façon à réduire l'influence de l'erreur de calcul. Comme pour $k = 2$, on introduit par exemple les variables auxiliaires

$$z_n^i = \nabla^i y_n h^{-i}, \quad i = 1, \ldots, k-1.$$

Essayons d'expliquer l'amélioration du schéma aux différences (5) lorsqu'on passe aux formules de calcul (7) du point de vue de la quantité d'information conservée. Dans le cas de la formule (5) sont implantées en mémoire, pour chaque n, les quantités y_{n-1}, y_n, sur lesquelles on va déterminer les valeurs y_j suivantes. Soit, pour fixer les idées, $1/2 \leqslant y_{n-1}$, $y_n \leqslant 1$ et $|y_n - y_{n-1}| \leqslant Mh$. Alors la cellule où est logée la valeur $y_{n-1} = 0, 1\alpha_2 \ldots \alpha_t$ est à $t-1$ bits $\alpha_2, \ldots, \alpha_t$ indépendants ; les bits du nombre $y_n = 0, 1\beta_2 \ldots \beta_t$ ne portent plus d'information nouvelle. La différence $y_n - y_{n-1} = \pm 0, 0\gamma_2 \ldots \gamma_t$ a un module inférieur à 2^{-l}, où l est le plus grand entier tel que $Mh < 2^{-l}$. C'est pourquoi

$$y_n - y_{n-1} = \pm 0, 0 \ldots 0\gamma_{l+1} \ldots \gamma_t,$$

et $t - l + 1$ bits (signe de différence et $\gamma_{l+1}, \ldots, \gamma_t$) suffisent pour se donner $y_n - y_{n-1}$, donc y_n. Evidemment $l \sim \log_2((Mh)^{-1})$, et, par conséquent, la quantité d'information indépendante qu'on a au total pour chaque n est d'environ $2t - \log_2((Mh)^{-1})$ bits.

Dans le calcul selon la formule (7), tous les bits occupés par les nombres y_n et z_n sont indépendants et l'information sur la solution est donnée par des signes binaires indépendants.

Le fait que dans le second procédé la quantité d'information indépendante est supérieure ne signifie pas pour autant qu'il est meilleur. Car l'information complémentaire peut s'avérer vide de toute signification, d'où l'impossibilité de préciser la solution. Montrons que l'information complémentaire a un sens dans le deuxième procédé.

Pour définir à $O(\varepsilon)$ près une solution d'une équation différentielle d'ordre 2 il faut se donner avec la même précision les valeurs de la solution et de sa dérivée en un point. Dans le cas d'une équation aux différences ces quantités sont y_n et $\frac{y_n - y_{n-1}}{h}$. En nous donnant, dans le premier procédé, les valeurs de y_n et y_{n-1} avec t signes binaires, nous déterminons $\frac{y_n - y_{n-1}}{h}$ avec une erreur en $O(2^{-t}h^{-1})$. L'information à notre disposition nous permet donc de trouver les valeurs suivantes de la solution du problème discret avec une précision de $O(2^{-t}h^{-1})$ (si tous les résultats ultérieurs sont strictement exacts). Dans le second procédé, nous avons les valeurs de y_n et $\frac{y_n - y_{n-1}}{h}$ à t bits et nous pouvons déterminer la solution du problème discret à $O(2^{-t})$ près. Ainsi, l'information complémentaire en question a bien un sens. A chaque pas de l'intégration numérique réelle les erreurs d'arrondi apportent une indétermination supplémentaire dans les valeurs du vecteur $\left(y_n, \frac{y_n - y_{n-1}}{h}\right)$, qui est en $O(2^{-t}/h)$ la première fois et en $O(2^{-t})$ la deuxième. Si bien que l'erreur de calcul totale sur la solution peut être de l'ordre de $O(2^{-t}/h^2)$ pour le premier procédé et de $O(2^{-t}/h)$ pour le second.

Indiquons une technique d'*évaluation pratique de l'erreur de calcul par changement d'échelle*. Soit à résoudre le problème de Cauchy

$$\frac{dy}{dx} = f(x, y), \quad y(0) = a,$$

par une méthode. Le changement de variables $x = \mu t$, $y = \lambda z$ ramène ce problème à

$$\frac{dz}{dt} = \frac{\mu}{\lambda} f(\mu t, \lambda z), \quad z(0) = \frac{a}{\lambda}.$$

Supposons que dans le premier cas on a intégré avec les pas h_i et intégrons numériquement le second problème par la même méthode mais avec les pas $h_i' = h_i/\mu$.

S'il n'y a pas d'arrondi, $y_i \equiv \lambda z_i$; si λ et μ ne sont ni l'un ni l'autre des puissances entières de deux, la différence entre les valeurs effectives y_i et λz_i donne généralement une idée de la grandeur de l'erreur de calcul. On prend par exemple $\mu = \sqrt{3}$, $\lambda = \sqrt{2}$.

Dans une intégration avec une recherche automatique du pas on fera attention à choisir la nouvelle valeur du paramètre ε et le pas de départ h_0 de sorte que $h_i' = h_i/\mu$ pour n'importe quel i (consulter pour les détails § 17, chap. III).

§ 13. Evaluation de l'erreur de calcul pour un problème aux limites résolu par la méthode du balayage

Nous avons promis, lorsque nous avons parlé des méthodes numériques de l'algèbre, de donner un exemple d'évaluation de l'erreur de calcul résultant de la résolution des systèmes algébriques par la méthode de Gauss. Faisons-le pour le système d'équations algébriques apparu lors de l'approximation du problème (1.3), (1.4).

Le système en question s'écrit

$$y_{j-1} - (2 + p_j h^2)\, y_j + y_{j+1} = f_j h^2,\ j = 1, \ldots, N-1,\ y_0 = a,\ y_N = b. \quad (1)$$

Dans l'algorithme d'élimination de Gauss appliqué aux inconnues dans l'ordre $y_0, \ldots, y_{N-1}$, le calcul s'effectue par

$$C_{n+1} = (2 + p_n h^2 - C_n)^{-1},\ n = 0, \ldots, N-2,\ C_0 = 0, \quad (2)$$

$$\varphi_{n+1} = C_{n+1}(\varphi_n - f_{n+1} h^2),\ n = 0, \ldots, N-2,\ \varphi_0 = a, \quad (3)$$

puis par

$$y_n = C_n y_{n+1} + \varphi_n,\ n = N-1, \ldots, 1. \quad (4)$$

Tout nombre z, résultat d'une opération arithmétique sur deux nombres en machine, est ensuite arrondi à un certain $z^* = z(1+\varepsilon)$; ceci étant, d'ordinaire $|\varepsilon| \leqslant 2^{-t}$, où t est la longueur des nombres en machine. Supposons que la dernière inégalité est toujours valable. Désignons par les lettres grecques des quantités de module inférieur ou égal à une constante absolue multipliée par 2^{-t} et par z^* une valeur calculée de z. Du moment que nous nous intéressons à l'erreur de calcul commise au cours de la résolution du système, admettons, par souci de simplicité, que les coefficients a, b, $2 + p_j h^2$, $f_j h^2$ sont donnés sans arrondi. Dans le cas contraire, les relations (2), (3), (4) se mettent sous la forme

$$C_{n+1}^* = (1 : (((2 + p_{n+1} h^2) - C_n^*)(1+\beta_n)))(1+\alpha_n), \quad (5)$$

$$\varphi_{n+1}^* = (C_{n+1}^*((\varphi_n^* - f_{n+1} h^2)(1+\gamma_n)))(1+\sigma_n), \quad (6)$$

$$y_n^* = ((C_n^* y_{n+1}^*)(1+\omega_n) + \varphi_n^*)(1+\delta_n). \quad (7)$$

Chaque parenthèse de la forme $(a \cdot b)$ avec $\cdot$ le symbole d'une opération arithmétique, est suivie d'un facteur $1+\varepsilon$ correspondant à cette opération avec arrondi. On récrit (5):

$$C_{n+1}^* = (2 + p_{n+1}^* h^2 - C_n^*)^{-1},$$

où

$$p_{n+1}^* = p_{n+1} + g_{n+1},\quad g_{n+1} = \frac{(2 + p_{n+1} h^2 - C_n^*)(\beta_n - \alpha_n)}{(1+\alpha_n) h^2}. \quad (8)$$

La relation (7) prend la forme

$$\frac{y_n^*}{1+\delta_n} = C_n^* y_{n+1}^* + \varphi_n^* + \omega_n C_n^* y_{n+1}^*$$

ou encore

$$y_n^* = C_n^* y_{n+1}^* + \varphi_n^{**}$$

avec $\varphi_n^{**} = \varphi_n^* + d_n$, où

$$d_n = \omega_n C_n^* y_{n+1}^* + \frac{y_n^* \delta_n}{1+\delta_n}.$$

Portons $\varphi_n^* = \varphi_n^{**} - d_n$, $\varphi_{n+1}^* = \varphi_{n+1}^{**} - d_{n+1}$ dans (6) écrit comme suit :

$$\frac{\varphi_{n+1}^*}{(1+\gamma_n)(1+\sigma_n)} = C_{n+1}^* (\varphi_n^* - f_{n+1} h^2);$$

il vient

$$\frac{\varphi_{n+1}^{**} - d_{n+1}}{(1+\gamma_n)(1+\sigma_n)} = C_{n+1}^* (\varphi_n^{**} - d_n - f_{n+1} h^2).$$

D'où

$$\varphi_{n+1}^{**} = C_{n+1}^* (\varphi_n^{**} - f_{n+1}^* h^2),$$

$$f_{n+1}^* - f_{n+1} = \varphi_{n+1} =$$

$$= \frac{1}{h^2}\left(d_n - (C_{n+1}^*)^{-1}\left(\frac{d_{n+1}}{(1+\gamma_n)(1+\sigma_n)} + \frac{\gamma_n + \sigma_n + \gamma_n \sigma_n}{(1+\gamma_n)(1+\sigma_n)}\,\varphi_{n+1}^{**}\right)\right). \tag{9}$$

Les égalités ci-dessus entraînent que les valeurs y_n^* calculées coïncident avec les valeurs obtenues lors du calcul sans arrondi du système

$$\begin{gathered} y_{j-1}^* - (2 + p_j^* h^2) y_j^* + y_{j+1}^* = f_j^* h^2, \quad j = 1, \ldots, N-1, \\ y_0^* = a, \quad y_N^* = b. \end{gathered} \tag{10}$$

L'évaluation de l'erreur de calcul s'est donc ramenée à l'évaluation du degré de proximité des solutions des systèmes (1) et (10) à coefficients voisins. Ces systèmes peuvent être écrits sous une forme plus commode

$$l_h(y_j) = \frac{y_{j+1} - 2y_j + y_{j-1}}{h^2} - p_j y_j = f_j, \tag{11}$$

$$l_h^*(y_j^*) = \frac{y_{j+1}^* - 2y_j^* + y_{j-1}^*}{h^2} - p_j^* y_j^* = f_j^*. \tag{12}$$

Limitons-nous au cas $p(x) \geqslant 0$ où l'estimation d'erreur s'obtient d'une manière particulièrement simple et supposons $|p_n h^2| \leqslant 2$ et $|f_n h^2| \leqslant \text{const}$. En retranchant (12) de (11) nous avons

$$l(R_j) = f_j - f_j^* + (p_j^* - p_j) y_j^* = v_j$$

donnant l'erreur $R_j = y_j - y_j^*$. Par suite du lemme du § 1 on a

$$\max_j |R_j| \leqslant \frac{1}{4} \max_j |v_j| \leqslant \frac{1}{4} (\max_j) |f_j - f_j^*| + (\max_j |p_j - p_j^*|) \max_j |y_j^*|). \tag{13}$$

Tout nombre non inférieur à 2^s (s entier) est arrondi à une valeur 2^s au moins et tout nombre non supérieur à 2^s à une valeur 2^s au plus. Evidemment $C_1 = \frac{1}{2 + p_1 h^2} \geqslant \frac{1}{4}$; si $\frac{1}{4} \leqslant C_n^* \leqslant 1$, tout ce que nous avons dit plus haut et la

formule de récurrence (2) entraînent

$$C_{n+1}^* = \frac{1}{2 - C_n^*} \leqslant \frac{1}{2-1} = 1, \quad C_{n+1}^* \geqslant \frac{1}{4 - C_n^*} \geqslant \frac{1}{4}.$$

Ainsi,

$$\max_n |C_n^*| \leqslant 1, \quad \max_n |(C_n^*)^{-1}| \leqslant 4.$$

Vu (8) nous avons l'estimation

$$\max_n |g_n| \leqslant C' 2^{-t} h^{-2};$$

il résulte de (9):

$$|q_n| \leqslant C'' 2^{-t} h^{-2} (|y_{n-1}^*| + |y_n^*| + |y_{n+1}^*|)$$

et, par conséquent,

$$\max_n |q_n| \leqslant C'' 2^{-t} h^{-2} \max_n |y_n^*|.$$

Portons ces estimations à droite dans (13); il vient

$$\max_j |y_j - y_j^*| \leqslant C^0 2^{-t} h^{-2} \max_j |y_j^*|,$$

ou encore

$$\| y - y^* \| \leqslant C^0 2^{-t} h^{-2} \| y^* \|. \tag{14}$$

Nous avons évalué la solution approchée du problème discret par sa norme. Faisons de même pour la solution exacte. La dernière inégalité implique

$$\| y^* \| \leqslant \| y \| + C^0 2^{-t} h^{-2} \| y^* \|,$$

d'où

$$(1 - C^0 2^{-t} h^{-2}) \| y^* \| \leqslant \| y \|, \quad \| y^* \| \leqslant \frac{\| y \|}{1 - C^0 2^{-t} h^{-2}},$$

lorsque $1 - C^0 2^{-t} h^2 > 0$; le report de cette estimation dans le second membre de (14) fournit

$$\| y - y^* \| \leqslant \frac{C^0 2^{-t} h^{-2}}{1 - C^0 2^{-t} h^{-2}} \| y \|.$$

Si l'on peut évaluer directement l'erreur de calcul à l'aide des coefficients du système initial, on pourra utiliser l'inégalité

$$\max_j |y_j| \leqslant \max (|y_0|, |y_N|) + \frac{1}{4} \max_j |f_j|.$$

La grosse erreur due au calcul avec arrondi semble attester la qualité médiocre de la méthode décrite pour le système (1). Or, une erreur de cet ordre de grandeur résulte déjà de l'enregistrement même du système. En général, le coefficient $2 + p_j h^2$ est logé en machine avec une erreur d'arrondi ε_j de l'ordre de 2^{-t}, i.e. ce coefficient s'écrit $s_j = 2 + p_j h^2 + \varepsilon_j$, ou encore

$$s_j = 2 + p_j^* h^2, \quad p_j^* = p_j + \varepsilon_j h^{-2}.$$

Ainsi, cet arrondi équivaut à perturber le coefficient p_j du système initial d'une quantité en $2^{-t} h^{-2}$. Conformément aux raisonnements du paragraphe 15 du chapitre précédent, cette circonstance est susceptible d'induire sur la solution du système (1) une perturbation du même ordre.

Pour réduire sensiblement l'erreur sur la solution de (1) il faut au moins se donner celui-ci sous une forme où l'erreur d'arrondi sur les coefficients du

système équivaut à des perturbations notablement moindres des coefficients du problème différentiel de départ. A cette fin, on passe par exemple au système

$$\frac{y_j - y_{j-1}}{h} = z_j, \quad \frac{z_{j+1} - z_j}{h} - p_j y_j = f_j.$$

Au lieu des relations de récurrence (2), (3) relativement aux coefficients C_k. φ_k on utilise respectivement les relations récurrentes (4.9), (4.10) sur α_k, β_k.

Bibliographie du chapitre IX

I [1, 4, 11, 17], II [1, 20, 28, 29, 31]; § 1: II [19], III [16]; § 2: II [28], III [8, 42]; § 3: II [6], III [4, 20]; § 4: III [43]; § 5 : III [40]; §§ 6, 7: III [1, 2, 3, 4, 20, 21, 36, 52]; § 8: II [1, 29, 32]; § 9: II [28], III [14, 28, 37, 39, 42]; § 10: II [29, 34], III [38, 51]; § 11: III [12, 49, 50]; § 12: II [1, 31], III [8]; § 13: III [8].

BIBLIOGRAPHIE

I. Ouvrages d'enseignement

1. Березин И. С., Жидков Н. П., Методы вычислений, т. 1, « Наука », 1966, т. 2, Физматгиз, 1962.
2. Воеводин В. В., Численные методы алгебры. Теория и алгорифмы, « Наука », 1966.
3. Гавурин М. К., Лекции по методам вычислений, « Наука », 1971.
4. Годунов С. К., Рябенький В. С., Разностные схемы (введение в теорию), « Наука », 1973.
5. Крылов А. Н., Лекции о приближенных вычислениях, Гостехиздат, М.-Л., 1950.
6. Крылов В. И., Приближенное вычисление интегралов, « Наука », 1967.
7. Крылов В. И., Бобков В. В., Монастырный П. Н., Вычислительные методы высшей математики, 1, «Вышэйшая школа», Минск, 1972.
8. Микеладзе Ш. Е., Численные методы математического анализа, Гостехиздат, 1953.
9. Фаддеев Д. К., Фаддеева В. Н., Вычислительные методы линейной алгебры, Физматгиз, 1960. Trad. and. Computational methods of linear algebra. W. H. Freeman and Cy., San Francisco, 1963.
10. Collatz L., Funktionalanalysis und numerische Mathematik, Springer Verlag, 1964.
11. Collatz L., Numerische Behandlung von Differentialgleichungen, Berlin (Springer), 1951.
12. Forsythe G. E., Moler C. B., Computer solution of linear algebraic systems, Englewood Cliffs (N. Y.), 1967.
13. Hamming R. W., Numerical methods for scientists and engineers, N. Y. [a. o.], McGraw-Hill, 1962.
14. Householder A. S., Principles of numerical analysis, McGraw-Hill, New York, 1953.
15. Lanczos C., Applied Analysis, Prentice Hall, Englewood Cliffs, N. Y., 1956.
16. Milne W. E., Numerical Calculus, Princeton U. P., 1949.
17. Milne W. E., Numerical solution of differential equations, New York-London, 1953.

II. Monographies et ouvrages divers

1. Бахвалов Н. С., Вычислительные методы решения обыкновенных дифференциальных уравнений, Ин-т кибернетики АН УССР, Киев, 1970.
2. Бусленко Н. П., Шрейдер Ю. А., Метод статистических испытаний (Монте-Карло) и его реализация на цифровых вычислительных машинах, Физматгиз, 1961.
3. Витушкин А. Г., Оценка сложности задачи табулирования, Физматгиз, 1959.
4. Воробьев Ю. В., Метод моментов в прикладной математике, Физматгиз, 1958.
5. Вычислительные методы линейной алгебры, под ред. Марчука Г. И., Изд-во СО АН СССР, Новосибирск, 1972.

6. Годунов С. К., Рябенький В. С., Введение в теорию разностных схем, Физматгиз, 1962.
7. Гончаров В. Л., Теория интерполирования и приближения функций, Гостехиздат, 1954.
8. Ермаков С. М., Метод Монте-Карло и смежные вопросы, « Наука », 1971.
9. Жилейкин Я. М., Халус Е., Стандартная программа для вычисления многократных интегралов с помощью квадратур Гаусса, вып. 39, ВЦ МГУ, 1970.
10. Загускин В. Л., Справочник по численным методам решения алгебраических и трансцендентных уравнений, Физматгиз, 1960.
11. Канторович Л. В., и Крылов В. И., Приближенные методы высшего анализа, Гостехиздат, М.-Л., 1949.
12. Копченова Н. В., Марон И. А., Вычислительная математика в примерах и задачах, « Наука », 1972.
13. Коробов Н. М., Теоретико-числовые методы в приближенном анализе, Физматгиз, 1963.
14. Крылов В. И., Шульгина Л. Т., Справочная книга по численному интегрированию, « Наука », 1966.
15. Лаврентьев М. М.; О некоторых некорректных задачах математической физики, Изд-во СО АН СССР, Новосибирск, 1962.
16. Линник Ю. В., Метод наименьших квадратов и основы математико-статистической теории обработки наблюдений, Физматгиз, 1962.
17. Марчук Г. И., Лебедев В. И., Численные методы в теории переноса нейтронов, Атомиздат, 1971.
18. Метод статистических испытаний (метод Монте-Карло), СМБ, Физматгиз, 1962.
19. Микеладзе Ш. Е., Новые методы интегрирования дифференциальных уравнений и их приложения к задачам теории упругости, Гостехиздат, М.-Л., 1951.
20. Михлин С. Г., Численная реализация вариационных методов, « Наука », 1966.
21. Моисеев Н. Н., Численные методы в теории оптимальных систем, « Наука », 1971.
22. Налимов В. В., Чернова Н. А., Статистические методы планирования экспериментов, « Наука », 1965.
23. Никольский С. М., Квадратура формулы, Физматгиз, 1958.
24. Процедуры численного интегрирования на языке Алгол, ВЦ МГУ, 1972.
25. Растригин Л. А., Статистические методы поиска, « Наука », 1968.
26. Ремез Е. Я., Общие вычислительные методы чебышевского приближения, Изд-во АН УССР, Киев, 1957.
27. Рябенький В. С., Филиппов А. Ф., Об устойчивости разностных уравнений, Гостехиздат, 1956.
28. Самарский А. А., Введение в теорию разностных схем, « Наука », 1971.
29. Шаманский В. Е., Методы численного решения краевых задач на ЭЦВМ, ч. I, Изд-во АН УССР, Киев, 1963; ч. II, « Наукова думка », Киев, 1966.
30. Ahlberg J. H., Nilson E. N., Walsh J. L., The theory of splines and their applications, Acad. Press, 1967.
31. Babuška I., Práger M., Vitásek E., Numerical processus in differential Equations, Interscience publishers, 1966.
32. Beckenbach E., Modern Mathematics for the Engineer, McGraw-Hill, 1956.
33. Brillouin L., La science et la théorie de l'information, Masson et Cie, Paris, 1959.
34. Collatz L., Eigenwert aufgaben mit technischen Anwendungen, Leipzig, 1963.
35. Dahlquist Y., Stability and error bounds in the numerical integration of ordinary differential equations, Uppsala, Almquist & Wiksells boktr 130 (1959), 5-92.

36. Guelfond A., Calcul des différences finies, Dunod, 1962.
37. Jánossy L., Theory and practice of the evaluation of measurements, Oxford, Clarendon press, 1965.
38. Jenkins G. M., Watts D., Spectral analysis and its applications, San Francisco, 1969.
39. Lattès R., J.-L. Lions, Méthode de quasi-réversibilité et applications, Paris, Dunod, 1967.
40. Morris W. T., Management science. A Bayesian introduction. Englewood Cliffs (N. Y.), 1968.
41. Morse Ph. M., Kimball G. E., Methods of operations research, London, Chapman Hall, 1951.
42. Ostrowski A. M., Solutions of equations and systems of equations. University of Basel. Basel Switzerland Academic Press. New York and London, 1960.
43. Platt W., Strategic intelligence production. Basic principles. N. Y., Praeger, 1957.
44. Stroud A. H., Secrest D., Gaussian Quadrature Formulas, Englewood Cliffs, N. Y., 1966.
45. Walsh J. L., Interpolation and approximation by rational functions in the complex domain. Am. Math. Soc., Coll. Publ., XX (1960).
46. Wilkinson J. H., The algebraic eifenvalue problem. Oxford, Clarendon Press, 1965.
47. Zoutendijk G., Methods of feasible directions, Elsevier Publishing Company, Amsterdam.

III. Articles

1. Абрамов А. А., Вариант метода прогонки, Ж. вычисл. матем. и матем. физ. 1, № 2 (1961), 349-351.
2. Абрамов А. А., О переносе граничных условий для систем обыкновенных дифференциальных уравнений (вариант метода прогонки), Ж. вычисл. матем. и матем. физ. 1, № 3 (1961), 542-543.
3. Абрамов А. А., О переносе условия ограниченности для некоторых систем обыкновенных линейных дифференциальных уравнений, Ж. вычисл. матем. и матем. физ. 1, № 4 (1961), 733-737.
4. Абрамов А. А., Андреев В. Б., О применении метода прогонки к нахождению периодических решений дифференциальных и разностных уравнений, Ж. вычисл. матем. и матем. физ. 3, № 2 (1963), 377-381.
5. Бабурин О. В., Лебедев В. И., О вычислении интегралов в смысле главного значения, весов и узлов квадратурных формул Гаусса, Ж. вычисл. матем. и матем. физ. 5, № 3 (1965), 454-462.
6. Бахвалов Н. С., К оценке ошибки при численном интегрировании дифференциальных уравнений экстраполяционным методом Адамса, ДАН СССР **104**, № 5 (1955), 683-686.
7. Бахвалов Н. С., Некоторые замечания к вопросу о численном интегрировании дифференциальных уравнений методом конечных разностей, ДАН СССР **104**, № 6 (1955), 805-808.
8. Бахвалов Н. С., О накоплении вычислительной погрешности при численном решении дифференциальных уравнений. Сб. « Вычисл. методы и программирование », вып. I, Изд-во МГУ, 1962, 47-68.
9. Бахвалов Н. С., Об определении начального шага и оценке главного члена погрешности при численном интегрировании с автоматическим выбором шага, Сб. « Вычисл. методы и программирование », вып. I, Изд-во МГУ, 1962, 69-79.
10. Бахвалов Н. С., Об оптимальных оценках скорости сходимости квадратурных процессов и методов интегрирования типа Монте-Карло на классах функций, Сб. « Числ. методы решения диффер. и интегр. уравнений и квадратурные формулы », « Наука », 1964, 5-63.

11. Бахвалов Н. С., Об алгоритмах выбора шага интегрирования, Сб. « Вычисл. методы и программирование », вып. V, Изд-во МГУ, 1966, 3-8.
12. Бахвалов Н. С., К оптимизации методов решения краевых задач при наличии пограничного слоя, Ж. вычисл. матем. и матем. физ. 9, № 4 (1969), 841-859.
13. Бахвалов Н. С., Об оптимальности линейных методов приближения операторов на выпуклых классах функций, Ж. вычисл. матем. и матем. физ. 11, № 4 (1971), 1014-1018.
14. Бахвалов Н. С., Об оптимизации методов решения обыкновенных дифференциальных уравнений с сильно осциллирующими решениями, Ж. вычисл. матем. и матем. физ. 11, № 5 (1971), 1318-1322.
15. Бродский М. Л., Асимптотические оценки погрешности при численном интегрировании систем обыкновенных дифференциальных уравнений разностными методами, ДАН СССР 93, № 3 (1953), 599-602.
16. Волков Е. А., Эффективные оценки погрешности разностных решений краевых задач для обыкновенного дифференциального уравнения. Тр. МИАН 112 (1971), стр. 141-151.
17. Валеев К. Г., Ракитский Ю. Р., Об отыскании разностными методами областей неустойчивости решений системы линейных дифференциальных уравнений, Сб. « Числ. методы решения диффер. и интегр. уравнений и квадратурные формулы », « Наука », 1964, 110-122.
18. Гайсарян С. С., О построении начальных условий при интегрировании систем обыкновенных дифференциальных уравнений методом Адамса, Сб. « Вычисл. методы и программирование », вып. III, Изд-во МГУ, 1965, 262-265.
19. Гельфанд И. М., Цейтлин М. Л., Принцип нелокального поиска в системах автоматической оптимизации, ДАН СССР 137, № 2 (1961), 295-298.
20. Годунов С. К., Метод ортогональной прогонки для решения систем разностных уравнений, Ж. вычисл. матем. и матем. физ. 2, № 6 (1962), 972-982.
21. Годунов С. К., О численном решении краевых задач для систем линейных обыкновенных дифференциальных уравнений, УМН 16, вып. 3 (1961), 171-174.
22. Гончар А. А., О скорости рациональной аппроксимации непрерывных функций с характерными особенностями, Матем. сб. 73, вып. 4 (1967), 630-638.
23. Горбунов А. Д., Шахов Ю. А., О приближенном решении задачи Коши для обыкновенных дифференциальных уравнений с наперед заданным числом верных знаков, I, Ж. вычисл. матем. и матем. физ. 3, № 2 (1963), 239-253; II, там же, 6, № 3 (1964), 426-433.
24. Диткин В. А., Люстерник Л. А., Об одном способе практического гармонического анализа на сфере, Сб. « Вычисл. матем. и вычисл. техника », № 1, Изд-во АН СССР, 1953, 3-13.
25. Долгополова Т. Ф., Иванов В. К., О численном дифференцировании, Ж. вычисл. матем. и матем. физ. 6, № 3 (1966), 570-576.
26. Дьяконов Е. Г., О построении итерационных методов на основе использования операторов, эквивалентных по спектру, Ж. вычисл. матем. и матем. физ. 6, № 1 (1966), 12-34.
27. Жилейкин Я. М., О квадратурных формулах на классах функций, Ж. вычисл. матем. и матем. физ. 8, № 3 (1968), 507-516.
28. Ильин А. М., Разностная схема для дифференциального уравнения с малым параметром при старшей производной, Матем. заметки 6, № 2 (1969), 237-248.
29. Канторович Л. В., Функциональный анализ и прикладная математика, УМН 3, вып. 6 (28) (1948), 89-181.
30. Ким Г., О решении некоторых систем линейных алгебраических уравнений высокого порядка, Сб. « Вычисл. методы и программирование », вып. 3, Изд-во МГУ, 1965, 134-146.

31. Колмогоров А. Н., Тихомиров В. М., ε-энтропия и ε-емкость множеств в функциональных пространствах, УМН 14, вып. 2 (1959), 3-86.
32. Крейн С. Г., Шаблицкая И. О., Об устойчивости разностных схем для задачи Коши, Ж. вычисл. матем. и матем. физ. 6, № 4 (1966), 648-664.
33. Морозов В. А., О восстановлении функций методом регуляризации, Ж. вычисл. матем. и матем. физ. 7, № 4 (1967), 874-881.
34. Морозов В. А., О задаче дифференцирования и некоторых алгоритмах приближения экспериментальной информации, Сб. « Вычисл. методы и программирование », вып. 14, Изд-во МГУ, 1970, 46-62.
35. Мысовских И. П., О кубатурных формулах для вычисления интегралов по поверхности сферы, Сиб. матем. ж. 5, № 3 (1964), 721-723.
36. Огнева В. В., Метод « прогонки » для решения разностных уравнений, Ж. вычисл. матем. и матем. физ. 7, № 4 (1967), 803-812.
37. Повзнер А. Я., Павлов Б. В., Об одном методе численного интегрирования систем обыкновенных дифференциальных уравнений, Ж. вычисл. матем. и матем. физ. 3, № 4 (1973), 256-259.
38. Приказчиков В. Г., Однородные разностные схемы высокого порядка точности для задачи Штурма — Лиувилля, Ж. вычисл. матем. и матем. физ. 9. № 2 (1969), 315-336.
39. Ракитский Ю. В., Методика последовательного увеличения шага численного интегрирования систем обыкновенных дифференциальных уравнений, ДАН СССР 207, № 4 (1972), 793-795.
40. Рябенький В. С., Необходимые и достаточные условия хорошей обусловленности краевых задач для систем обыкновенных разностных уравнений, Ж. вычисл. матем. и матем. физ. 4, № 2 (1964), 242-255.
41. Салихов Н. П., О полярных методах решения задачи Коши для систем обыкновенных дифференциальных уравнений, Ж. вычисл. матем. и матем. физ. 2, № 4 (1962), 515-528.
42. Самарский А. А., Тихонов А. Н., Об однородных разностных схемах, Ж. вычисл. матем. и матем. физ. 1, № 1 (1961), 5-63.
43. Соболев С. Л., Некоторые замечания о численном решении интегральных уравнений, Изв. АН СССР, сер. матем. 20, № 4 (1956), 413-436.
44. Соболев С. Л., О формулах механических кубатур на поверхности сферы, Сиб. матем. ж. 3, № 5 (1962), 769-796.
45. Тихонов А. Н., О решении некорректно поставленных задач и методе регуляризации, ДАН СССР 151, № 3 (1963), 501-504.
46. Тихонов А. Н., Об устойчивости алгоритмов для решения вырожденных систем линейных алгебраических уравнений, Ж. вычисл. матем. и матем. физ. 5, № 4 (1965), 718-722.
47. Тихонов А. Н., О некорректных задачах оптимального планирования, Ж. вычисл. матем. и матем. физ, 6, № 1 (1966), 81-89.
48. Тихонов А. Н., Горбунов А. Д., Асимптотические разложения погрешности разностного метода решения задачи Коши для системы обыкновенных дифференциальных уравнений, Ж. вычисл. матем. и матем. физ. 2, № 4 (1962), 536-548.
49. Тихонов А. Н., Горбунов А. Д., Оценки погрешности метода типа Рунге — Кутта и выбор оптимальных сеток, Ж. вычисл. матем. и матем. физ. 4, № 2 (1964), 232-241.
50. Тихонов А. Н., Горбунов А. Д., Гайсарян С. С., Описание алгоритма построения оптимальных сеток при решении задачи Коши для обыкновенных дифференциальных уравнений методом Рунге — Кутта, Сб. « Вычислительные методы и программирование », вып. 6, Изд-во МГУ, 1966, 17-20.
51. Тихонов А. Н., Самарский А. А., Разностная задача Штурма — Лиувилля, Ж. вычисл. матем. и матем. физ. 1, № 5 (1961), 784-805.
52. Шапкин А. Ф., Замыкание двух вычислительных алгоритмов, основанных на идее ортогонализации, Ж. вычисл. матем. и матем. физ. 7, № 2 (1967), 411-416.

53. Bulirsch R., Bemerkungen zur Romberg-Integration, Numer. Math. 6, № 11 (1964), 6-16.
54. Butcher I. G., A Modified Multistep method for the numerical integration of ordinary differential equations, J. Assoc. Comput. mach. vol. 12, № 1 (1965), 124-135.
55. Dahlquist Y., Convergence and stability in the numerical integration of ordinary differential equations, Math. Skand., vol. 4, № 1 (1956), 33-53.
56. Distefano G. P., Causes of instabilities in numerical integration techniques, Internat. J. Comput. Math. vol. 2, № 2 (1968), 123-142.
57. Filippi S., Kraske F., Neue verallgemeinerte Multistep Methoden., ZAMM 48, № 8 (1968), Sonderh, 65-66.
58. Fyfe D. J., Economic evaluation of Runge-Kutta formulae, Math. Comp. 20, 95 (1966), 392-398.
59. Gragg W. B., Stetter H. I., Generalized Multistep Predictor Corrector Methods, J. Assoc. Comput. Mach., vol 11, № 2 (1964), 188-209.
60. Grane R. L., Klopenstein. A Predictor-Corrector Algorithm with an Increased Range of Absolute Stability, J. Assoc. Comput. Mach., vol. 12, № 2 (1965), 227-242.
61. Haber S., Numerical evaluation of multiple integral, SIAM Rev., vol. 12, № 4 (1970), 481-526.
62. Liniger W., A criterion for a stability of linear multistep integration formulae, Computing 3, № 3 (1968), 280-285.
63. Reinsch R., Smoothing by Spline Functions, Numer. Mat. 10 (1967), 177-183.

INDEX

Accélération de la convergence de la méthode de Monte-Carlo 296, 299
Application contractante 392
Approximation
  d'une condition aux limites 521
  d'une équation différentielle par un schéma aux différences 454, 518

Borne supérieure d'une erreur 24

Chiffre(s)
  exact 21
  significatifs 21
Coefficients de Fourier discrets 211
Colonne de contrôle d'un système 316
Conditionnement
  de la matrice d'un système 375
  d'un système 375
Condition α 464
Conditions de convergence dans la résolution du problème de Cauchy par les méthodes des différences finies 464, 504
  initiales pour un schéma aux différences finies 492
Contrôle de l'erreur par pas 442

Dégagement de la fonction de poids 179
Dérivation numérique 88
Dérivée d'un opérateur 396
Différences
  ascendantes 61
  centrales 61
  descendantes 61
  divisées 37
  finies 61
  d'ordre supérieur 61
  du premier ordre 61
Dissipation 414
Domaine de convergence d'une méthode 391

Elévation de l'ordre de précision d'un schéma aux différences 518
Elimination des racines 405
Entropie 252
ε-entropie 255
Equation caractéristique 50
  d'un schéma aux différences 459
Equation différentielle donnée sous forme implicite 495
Equation aux différences linéaire à coefficients constants 49, 55
  d'ordre k 46
Equation aux différences majorante 588
Equations aux différences finies 46
Erreur
  absolue 21
  d'approximation d'une équation différentielle par un schéma aux différences 454
  de calcul 17
  — dans la dérivation numérique 91
  — dans l'intégration numérique 589
  — dans l'interpolation numérique 75
  — dans la résolution des problèmes d'algèbre 317, 330, 593
  — — du problème de Cauchy 443, 465, 587
  — — des problèmes aux limites 528, 550, 592, 593
  de formulation 20
  dans une formule de quadrature sur une classe de fonctions 102, 135, 281
  inhérente 17, 20
  de méthode 17
  — sur une classe de problèmes 60, 135
  — sur un pas 436, 470
  relative 21
Espace métrique 30
Espace vectoriel 30
  normé 30
  strictement normé 31
Estimation linéaire d'une erreur 25
Estimations garanties de l'erreur sur une classe de fonctions 289

Evaluation de l'erreur de calcul par changement d'échelle 317, 592
Evaluation pratique
de l'erreur dans des formules de quadrature 117, 161
de l'erreur d'interpolation 41, 66, 507-508
de l'erreur résultant de la résolution d'équations différentielles 473, 519, 592
— — — de systèmes linéaires 317, 336
Extrapolation 32

Fermeture
d'un algorithme de calcul 536
irrégulière 537
de la méthode du balayage 538, 555
régulière 537
Fonction(s)
économique 420
élémentaires 239
de Green discrète 522
majorante 522
de poids 95
Fonction-spline 247
Formule(s)
d'Adams 450
de Filon 123
de Gregory 160-161
d'Hermite 115
de Lobatto 112
de Markov 112
de Newton 65
des rectangles 94, 97
— à point multiple 98
de Romberg 168
de Simpson 99
des trapèzes 94, 98
Formules de dérivation numérique 88
unilatérales 91
Formules d'interpolation 449
de Bessel 68
d'Everett 71
de Gauss 67
— pour l'interpolation en arrière 67
— — en avant 67
de Lagrange 36
de Newton par différences divisées 39
— pour l'interpolation en arrière 65
— — en avant 65
— pour des points équidistants 65
Formules de quadrature
composites 128
d'Euler 159
de Filon 123
de Gauss 106
généralisées 128
de Gregory 160-161
d'Hermite 115
de Lobatto 112
de Markov 112
de Newton-Cotes 94
optimales 130, 135, 136
aux points aléatoires 299
des rectangles 94, 97
de Romberg 169
de Simpson 99
de Tchebycheff 101
des trapèzes 94, 98

Inégalité de Tchebycheff 290
Intégration
des équations différentielles dans un grand intervalle 447, 487, 490, 597
des fonctions oscillantes 122
de systèmes d'équations 483
Interpolation 32, 34
basée sur des points multiples 42
inverse 77
linéaire 69
quadratique 70
trigonométrique 211
Inversion de matrices 319

Machine
travaillant en virgule fixe 20
— — flottante 20
Matrice
$(2m_1 + 1)$-diagonale 533
d'orthogonalisation 80
Meilleur approximant 201
Meilleure approximation uniforme 215
Méthode(s)
d'Adams 451
d'Adams-Störmer 500, 502
d'amélioration de la précision par des différences d'ordre supérieur 520
asymptotiquement optimale 142
du balayage 530, 551
— orthogonal différentiel 555, 557, 559
— par remontée 531
— en sens direct 531
bilatérales d'intégration numérique 508
— d'interpolation 500
— de résolution des équations différentielles 511
de la boule pesante 416
de Butcher 457
des coefficients indéterminés 34, 260, 454, 500
de descente 405

Méthode(s)
de descente séquentielle 352, 405, 408
— —, optimale 408
des différences finies 449
d'élimination de Gauss 312
d'encadrement 404, 560-561
d'Euler 434, 474
explicite 449, 500
d'extrapolation 449, 500
de Gauss 312
— avec choix du pivot 316
— directe 315
— par remontée 315, 531
de Gerschgorin 522
du gradient 355, 408
— conjugué 358
implicite 449, 494, 500
d'intégration numérique à pas liés 444
— — — séparés 444
d'interpolation 449, 500
itérative de construction du polynôme de meilleure approximation uniforme 225
— avec utilisation des opérateurs spectralement équivalents 371
itérative simple 322, 391
de Markov 112
de Milne 474
des moindres carrés 261
de Monte-Carlo 289
— pour des systèmes d'équations linéaires 364
de Newton pour des systèmes d'équations non linéaires 396
optimale 60, 131
d'orthogonalisation 320
des paraboles 405
de la plus grande pente 355, 408
probabiliste 302
de la racine carrée 317
des ravins 412
de régularisation 263, 379
de relaxation 354
des rotations de recherche des valeurs propres 385
de Runge-Kutta 434
de la sécante 401
de Seidel 349, 394
de stationnarisation 414
de surrelaxation 355
télescopique 223
du tir 528, 560-561

Nombre
grand 21
très grand 21
Nombre de conditionnement
d'une matrice 375
d'un problème 547
d'un système 375
Normalisation d'un schéma aux différences finies 456
Normes 30
associées 322
équivalentes 325
des vecteurs et des matrices 322

Objectif 420
Opérateur différence finie 63
Optimisation
de l'estimation d'une erreur d'interpolation 60
des méthodes 60, 131
de la répartition des points d'intégration 142, 145, 582
de la vitesse de convergence des processus itératifs 339
Ordre d'une méthode 395
Ordre de grandeur
de l'approximation 455
de l'erreur dans la méthode de Runge-Kutta 436

Paramétrisation d'un problème 422, 561
Partie principale de l'erreur
dans des formules de quadrature 154
dans l'intégration numérique d'équations différentielles 448, 469, 519
Pas d'un tableau 61
Point du spectre d'une famille de matrices 330
Points d'interpolation 34
Points du support de Tchebycheff 217
Polynôme(s)
d'Hermite 85
de Jacobi 84
de Laguerre 85
de Legendre 84
de meilleure approximation uniforme 216
orthogonaux 80, 84
qui s'écartent de zéro le moins possible 57
de Tchebycheff 55
— de deuxième espèce 85
— de première espèce 85
Problème avec couche limite 567, 585
Problème aux limites bien conditionné (bien posé) 547
Problème de valeurs propres
global 380
local 380
Problèmes mal posés 377

Procédé
d'Aitken 40
δ^2 d'accélération de la convergence 338
Processus itératif
linéaire optimal 346
à pas liés 343
— séparés 343
Produit direct des formules d'intégration, d'interpolation, de dérivation 277, 278, 308
Programmes usuels d'intégration numérique 185, 192

Rayon spectral
d'une famille de matrices 331
d'une matrice 331
Réduction de la dimension
des problèmes d'optimisation 410
des systèmes non linéaires 413
Règle de Runge 162, 473, 519
Remplissage d'un tableau 72
Reste de la formule de Lagrange 36

Schéma(s)
aux différences finies 449
explicite 449, 494, 500
homogènes 573
implicite 449, 494, 500
Série de Fourier discrète, finie 211
Sous-espace vectoriel 30
Spectre d'une famille de matrices 330
Symétrisation d'un système d'équations 340, 348, 379
Système(s)
bien conditionnés 377
d'éléments linéairement dépendants 30
— — indépendants 30, 78
— orthonormé 80
fondamental de solutions 53
mal conditionnés 377
orthogonal 78

Tableau de différences divisées d'une fonction 38
Théorème
d'alternance de Tchebycheff 216
sur l'évaluation d'erreur 465
de la partie principale de l'erreur 470
de Vallée Poussin 216
Transformation de Fourier
discrète 208
rapide 212

Variation du pas de l'intégration numérique 191, 496

A NOS LECTEURS

Les Editions Mir vous seraient très reconnaissantes de bien vouloir leur communiquer votre opinion sur le contenu de ce livre, sa traduction et présentation, ainsi que toute autre suggestion.

Ecrire à l'adresse: 1er Rijski péréoulok, 2,
Moscou, I-278, U.R.S.S.

Imprimé en Union Soviétique